CONTEXTUAL INFORMATION FOR APPLICATIONS IN VIDEO SURVEILLANCE

A Dissertation Presented to
the Faculty of the Department of Computer Science
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

By
Li Wei
December 2016

CONTEXTUAL INFORMATION FOR APPLICATIONS IN VIDEO SURVEILLANCE

Li Wei

APPROVED:

Dr. Shishir K. Shah, Chairman
Dept. of Computer Science

Dr. Jaspal Subhlok
Dept. of Computer Science

Dr. Edgar Gabriel
Dept. of Computer Science

Dr. Saurabh Prasad
Dept. of Electrical & Computer Engineering

Dean, College of Natural Sciences and Mathematics

Acknowledgements

I would love to thank Dr. Shishir K. Shah for being an excellent advisor and mentor for my research and life for the past five years in the University of Houston. I always consider myself lucky as his Ph.D. student for the enormous supports and guides he gave me in my path of research.

I also want to thank all the members of Quantitative Imaging and Computational Biology Laboratory for all the help and inspiration conversations that lead to many research ideas and publications.

Finally, I want to thank my parent for their supports during my study. I want to give my greatest thank to my wife, Yajun, without her love and support, completing this Pd. D. degree would not be possible.

CONTEXTUAL INFORMATION FOR APPLICATIONS IN VIDEO SURVEILLANCE

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Li Wei

December 2016

Abstract

With a growing network of cameras being used for security applications, video-based monitoring relying on human operators is ineffective and lacking in reliability and scalability. In this thesis, I present automatic solutions that enable monitoring of humans in videos, such as identifying same individuals across different cameras (human re-identification) and recognizing human activities.

Analyzing videos using only individual-based features can be very challenging because of the significant appearance and motion variance due to the changing viewpoints, different lighting conditions, and occlusions. Motivated by the fact that people often form groups, it is feasible to model the interaction among group members to disambiguate the individual features in video analysis tasks. This thesis introduces features that leverage the human group as contextual information and demonstrates its performance for the tasks of human re-identification and activity recognition. Two descriptors are introduced for human re-identification. The Subject Centric Group (SCG) feature captures a persons group appearance and shape information using the estimate of persons' positions in 3D space. The metric is designed to consider both human appearance and group similarity. The Spatial Appearance Group (SAG) feature extracts group appearance and shape information directly from video frames. A random-forest model is trained to predict the group's similarity score. For human activity recognition, I propose context features along with a deep model to recognize the individual subjects activity in videos of real-world scenes. Besides the motion features of the person, I also utilize group context information and scene context information to improve the recognition performance.

This thesis demonstrates the application of proposed features in both problems. Our experiments show that proposed features can reach state-of-the-art accuracy on challenging re-identification datasets that represent real-world scenario, and can also outperform state-of-the art human activity recognition methods on 5-activities and 6-activities versions of the Collective Activities Dataset.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With a growing network of cameras being deployed in public places such as the college campuses, airports, and office buildings, a huge amount of video data is generated by these camera networks. These data are usually monitored by a human operator or utilized for forensic purpose [6]. Manual video monitoring that relies on a human operator is ineffective and lacking in reliability and scalability [41, 82], seriously reducing the effectiveness of surveillance. Therefore, an automatic solution to analyze and extract useful information from the massive videos has received increasing attention from the computer-vision community. Video analysis can enable long-term activity and behavior characterization of people in the scene, and such analysis is required for high-level surveillance tasks such as suspicious activity detection or undesirable event prediction for timely alerts [30]. Here are a few application areas that described in [69] that show the potential applications of this topic:

1. **Security and Surveillance:** Traditionally, most surveillance systems nowadays rely on a camera network that is monitored by human operators. However, With a growing network of cameras being used for security applications, manual re-identification that relies on a human operator is ineffective and lacking in reliability and scalability [41, 82]. The goal of this area is to develop vision-based solutions to these tasks that can assist even replace a human operator.

2. **Behavioral Biometrics:** The study of biometric targets developing algorithms for uniquely recognizing humans based on physical or behavioral cues. Fingerprint, face, and iris are widely used physiological biometrics in traditional approaches, which main relies on physical attributes for recognition. The limitation of these approaches is that they require the cooperation from the subject in order to collect the biometric. Behavior is as useful a cue to recognize humans as their physical attributes. The advantage of behavior biometric based approach is that subject cooperation is not necessary and it can proceed without interrupting or interfering with the subjects activity. Currently, the most-promising example of behavioral biometrics is human gait [64].

3. **Content-Based Video Analysis:** With the spreading of digital technologies, videos are more available to everyone and becoming an important material to record people's live and memory. The fast growth of video websites, such as YouTube, make it necessary to develop a technology to generate the semantic index that based on video content for better video summarization

and searching. The recent development of deep learning technologies has already shown the promising results for video labeling using deep convolution neural network [38].

4. **Animation and Synthesis:** The technology of synthesizing realistic human behavior is widely required by both gaming and movie industries. At the group/crowd level, the state-of-the-art crowd simulation approaches usually use a data-driven approach that learns the group behavior directly from videos [47]. At the individual level, it traditionally relied more on human animators to provide detail animations. Recently, capturing high-quality human behavior is possible due to the development of computer vision technology. Therefore, more data-driven approaches are developed, such as facial animation [76], upper body [48], and fingers [37].

The main research direction this thesis is exploring how contextual information in the video can help video analysis tasks. We propose contextual features that capture the human groups and scene as contextual information. We demonstrate the power of contextual information using the human re-identification and human activity recognition video analytic application. Our experiments show that proposed features can reach state-of-the-art accuracy on challenging reidentification datasets that represent real-world scenario, and can also outperform state-of-theart human activity recognition methods on 5-activities and 6-activities versions of the Collective Activities Dataset.

1.1 Challenges

The challenge of video analysis comes for multiple reasons. Firstly, it is difficult to design a robust and discriminative visual or motion features that need to be extracted from video data which is captured in unconstrained environments, where environmental variations, such as illuminations/viewpoints changing, as well as scene occlusions and clutter can happen to the objects in the videos.

Secondly, human-related information makes a great part of the information that extracted from videos. However, analyzing video that contains humans requires person detection and tracking to capture the input images of target subject for accurate feature extraction. Person detection and tracking are difficult problems with their own hurdles. There are a significant amount of works be published in person detection and tracking over the past two decades, but sustained tracking under varying observation environment remains an open problem. All these factors may lead to incorrect detections and faulty trajectory estimation, which introduce errors in the feature extraction that directly affect the video analysis task.

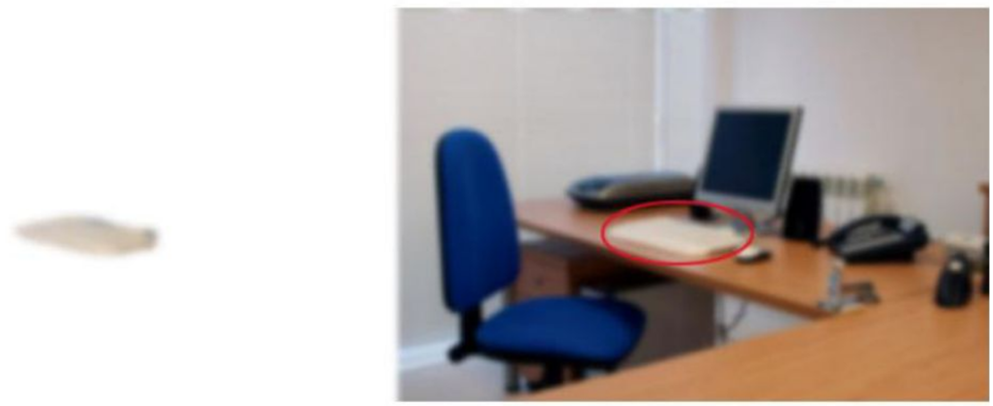
Thirdly, some applications try to recognize the human behavior by analyzing video data. However, human activity recognition suffers Intra- and inter-class variations as many human activities variations in performance. For instance, walking activity may differ in speed, stride length, and personal anthropometric. Also, similar observations can be made for different actions.

Figure 1.1: The different appearances of the same person in two camera views.

In summary, video analysis is a broad and difficult problem with numerous open issues. However, this thesis discovers that those challenges came be overcome by contextual information, which is the information provided by other objects, humans, and environments in the video scene and able to help to disambiguate video analyze tasks.

1.2 Motivations

In recent years, researchers in both human-vision and computer-vision communities have been greatly interested in the ability of contextual information in improving vision analysis tasks, such as object detecting and object recognition. The motivation of utilizing contextual information is attempting to enhance the current vision analysis technologies by utilizing the information from outside of the objects, such as scenes, group information, and meta data [54]. An example is shown in Figure 1.2. It's very hard to recognize the object by looking the object only in Fig 1.2(a); with context information provided in Fig 1.2(b), it provides rich contextual information, such as co-occurrence, spatial relationship, and make it very easy to recognize the keyboard object.



(a) (b)

Figure 1.2: A example showing the contextual information in object recognition. (a) A image of keyboard. (b) A keyboard displaying in the scene of context.

In this thesis, I explore the possible contextual information that helps improve the accuracy of video-analysis tasks. As people often stay/travel with each other, I found that the human group as contextual information is very helpful in person re-identification task, which associates the same person across different cameras. The group information that contains group structure and appearance information, by using which as additional evidence can leads to a better performance for re-identification tasks.

I also discover that the contextual information can also be helpful in video-based human activity recognition task. Group-contextual information for activity recognition captures the group structure and interaction information, which is very useful to recognize collective activities such as Queueing and Talking. The humans in these activities often form certain group shape and interaction pattern. I also discover scene information can help because certain activities are more likely to happen in certain locations. We demonstrate that by combining these contextual information, the accuracy of human activity recognition is greatly improved.

1.3 Contributions

This thesis presents the approaches of using human groups as contextual information and boosts the accuracy of video analysis tasks, we demonstrate proposed contextual feature in person re-identification and human activity recognition task.

Two types of contextual features is introduced for person re-identification task: subject centric group (SCG) feature and Spatial Appearance Group (SAG) feature. SCG feature captures the group appearance and shapes in the video. SAG also captures the group appearance/shape information, and it also enables us to train machine-learning models to learn the shape transformation between camera views. For human activity recognition, we introduce group contextual feature to capture and group shape and motion interaction among group members and a scene contextual features that encode the environment information.

To list my technical contribution:

1. To capture the human groups as contextual information, I propose Subject-Centric Group (SCG) feature [77] that encodes the person's profiles within the group, including in-group-position and co-travelers' appearance features. We also introduce the metrics for computing the distance between SCG features.

2. I propose Spatial-Appearance Group (SAG) feature [78] that encodes group appearance and shape in a fix length vector. A machines learning model is trained to learn the transformation of SAG between cameras. We propose a method to compute the similarity score given two observed persons in two cameras.

3. We introduce group contextual feature that captures the person's interaction in the video scene, and a scene contextual feature that encodes the environment person placed, as well as a pipeline that trains deep model which jointly consider both low-level individual motion feature and contextual features to recognize human activities in the video [79].

Experiments are performed to demonstrate that introduced Subject-Centric Group (SCG) features performs better than appearance-based approaches, and also out-performs other context-based re-identification methods. The experiments also show the Spatial-Appearance Group (SAG) slightly performs better than SCG on the testing database. The human activity recognition is evaluated using challenge video datasets that represent the real-life scenarios; the experiments show our deep model using contextual feature can outperform the state-of-the-art approaches.

1.4 Thesis Overview

The organization of the remainder of this dissertation is as follows. I begin in Chapter 2 by presenting a literature review of existing approaches for video analysis and human re-identification. Chapter 3 describes Subject-Centric Group (SCG) features and the designed metric for computing feature distances for re-identification. Chapter 4 introduces Spatial-Appearance Group (SAG) feature which captures both spatial and appearance information of the groups. Chapter 5 introduces how contextual features can be applied to human activity recognition tasks. I propose features that capture both human interactions and scene environment, as well as a deep model that recognize the human activities in video.

Chapter 2

Background

2.1 Video Analysis

In computer vision, the task of video analytics is automatically analyzing video to extract useful information, such as to detect and determine temporal and spatial events. As there are a great number of video cameras deployed in the scene such as campus, airports, and cities, computer vision communities have been drawing attention to developing the solution that can automatically monitor the video data from a lot of cameras. Summarized by [42], the video analysis framework, can include the following steps: motion/object detection, object classification, object tracking, behavior and activity analysis and understanding, person identification. The motion detection [5] aims to separate the regions corresponding to moving objects to the rest of the images. The subsequent processes such as object tracking and behavior analysis are greatly dependent on it. The object tracking [85, 4] aims to locate the moving object in am image sequence. Motion tracking usually has considerable interaction with motion detection during processing. In many applications, such as video surveillance and event detection, the major source of information comes from the human or human-controlled objects appearing in the videos. Understanding behavior [7, 66] involves analysis and recognition of motion patterns and the high-level description of actions and interactions among objects. Locating person of interest is always very important in video surveillance. Human face [70] and gait [72] are now used as main biometric features that can be used for personal identification. The above-discussed researches are mainly focused on single camera videos. However, in reality, most monitoring systems have multiple cameras, and it brings the problem of associating the objects among different videos. The task of associating persons between non-overlapping camera views is human Re-identification, and our current works focus on this topic. The related work of human re-identification will be given at Sec. 2.3.

2.2 Contextual Information

Humans have a great ability to detecting and recognizing objects, and performing a board range of visual tasks in a wide variety of situations, even with the considerable amount of clutter, occlusions, and illumination changes [54]. In human vision literature, many researchers [8, 9] suggest that contextual information plays a critical role in the human's ability to detect/recognize objects. The contextual information in object detection/recognition can come in two levels: semantic

[16, 17] encodes the co-vary between the recognition targets and scene or other objects in the scenes; spatial relations [19, 34] encodes the relatively locations information among recognition targets and other objects in the scene.

Motivated by the research of contextual information in human vision communities, there has been a surge of interest in context modeling for computer vision applications. The context has been an important information in image based object recognition task. Rabinovich and Belongie [62] introduce a classification of contextual models for object recognition task. Their models contain a model with contextual inference based on a statistical summary of the scene, and another model representing the context regarding the relationship among objects in the image. Peter et al. [11] propose a statical model that represent the co-occurrence relationship between the objects in the visual scene. The context information has also been used to analysis the human activities.

2.3 Human Re-Identification

In general, vision based person re-identification algorithm includes the representation step and matching step [26]. The representation step exploits low-level features such as color [53], texture [86] or their combinations [87] if single images are provided; there are also approaches that exploit spatio-temporal features [73, 24], accumulated appearance variability [29], and gaits [31, 55] if video or multiple frames are provided. These representations provide a reasonable level of inter-person discrimination as well as inter-camera invariance. These features can be further encoded into fixed-length vectors in forms of histogram, covariances and Fisher vectors. After the representation is obtained, a designed distance metric or model-based matching algorithms, such as support vector ranking [61], are used for re-identification. In the case of distance metrics, Euclidian distance [52] and Bhattacharyya distance [21] are used to computed the dissimilarity/distance between two samples. In the case of model-based matching algorithms, Li et al. [49] train random forest classifier based on the annotated data to calculate the similarity score. For a more comprehensive survey of re-identification approaches, please refer to the recent study found in [71, 27].

2.4 Contextual Information In Re-identification

The contextual information has been explored to improve re-identification in recent works. [88, 10] are the methods that most similar to ours. Zheng *et al.* [88] proposed a method to associate groups of peoples in non-overlapping camera views. Their approach explores group information as the contextual cue for reducing the ambiguity in person re-identification if a person appears in the group. They propose a rotation invariant descriptor named Center Rectangular Ring Ratio-Occurrence Descriptor (CRRRO) to handle the position change and camera viewpoint change. This approach addresses single shot re-identification and cannot easily extend to multiple shots scenarios. One limitation of the approach [88] is that it requires manually selected person group images as input, however, selecting group image itself is time-consuming because finding the people and detecting the groups manually in a large video datasets is quite tedious and requires expertise. The method introduced in this thesis detects people groups automatically by clustering the person trajectories, and the introduced person-group feature that is also robust to person position and camera viewpoint changes. Cai et al. [10] compute relative appearance context model of groups to decrease the ambiguities in individual appearance matching. Different to [88], Cai *et al.* use a relaxed definition of the group named neighboring set, which is a set of people that enter/exit at similar locations within a time frame. The groups under this definition have weak social connections. Therefore, the assumption that the same set of people will re-appear in different camera views is not very strong. Cai *et al.* [10] also assumes that appearance difference between a pair of people is similar across cameras. However, this assumption is also weak because the person appearance, as well as the difference of appearance, would significantly change due to the background, illumination and camera setting changes.

2.5 Human Activity Recognition

As it summarized in [39], a system of video-based human activity recognition comes in multiple levels. The low-level system focuses on pre-processing steps such as object segmentation, feature extraction, and representation, and action detection and classification algorithms. The mid-level system is human activity recognition system, which solves the problem of single person activity recognition, multiple people interaction and crowd behavior recognition, and abnormal activity detection. The high-level system supports many applications, such as surveillance environments and entertainment environments.

This thesis introduces the approaches can apply to mid-level human activity recognition systems, which require the pre-processing results at the low-level system. Many previous works focus on detecting people in a single image. David [51] introduce Scale-invariant feature transform (SIFT) that describe the local features in images, and SIFT feature is widely applied for detecting humans. Mikolajczyk et al. [57] improve the idea of SIFT by introducing a probabilistic model that detect human using assemblable individual parts, this work can detect the humans of various poses. Human tracking in the video also becomes a hot topic in recent years. A common human video-based human tracking method is a bottom-up approach, which means taking the human-detection results of every single frame and utilize particle filtering to obtain a smooth tracker. Okuna et al. [60] propose a method to compute tracker that initialize by detector output, and track by a practical filtering based on color information. Xu et al. [83] utilize a combine of pedestrian detection (using Histogram of Oriented Gradients - Support Vector Machine classifier) and particle filter to get pedestrian trajectory in the videos. Previous works have introduced various low-level features to describe the observed human action. Schuldt et al. [65] proposed a local space-time feature to represent the human movement observed in a video and integrated such representations with SVM classification schemes for recognition. Laptev et al. [45] proposed space-time feature point (STIP) and spatio-temporal bag-of-features as the descriptor for human motion. Tran et al. [67] presented a framework for human action recognition based on modeling the motion of human body parts. They utilized a descriptor that combines both local and global representations of human action, encoding the motion information and being robust to local appearance changes.

At mid-level, many previous works consider space-time information to recognize the activities. Ke et al. [40] introduce Spatio-temporal volume (SVT) to capture continuous human actions by concatenating image frames along the time axis. Kumari et al. [43] propose a Discrete Fourier Transformation (DFT) based methods that represent the human activity in the frequency domain. Both SVT and DFT are global features that consider the whole image, therefore, limited to occlusion and viewpoint change. Many previous works introduce local descriptors such as SIFT and HOG to capture the characteristics of image patch along the time. Jia et al. [36] build Local Spatio-Temporal Discriminant Embedding that projects a sequence of human silhouette into the embedding space as a representation and train the model to recognize human activity.

This thesis builds the representation of individual motion information based on STIP feature similar to [45], but combines the rich context information that extracted from the video. The introduces method can capture the extensive information about people motion and interactions; scale to recognize the activity of each individual in the scene, and improve the accuracy of the overall activity recognition task.

2.6 Contextual Information In Activity Recognition

Context information is widely utilized in recognizing the people group activity. Many approaches integrate contextual information by proposing new feature descriptors extracted from an individual and its surrounding area. Lan et al. [44] proposed Action Context (AC) descriptor capturing the activity of the focal person and the behavior of other persons nearby. The AC descriptor is concatenating the focal person action probability vector with context action vectors that capture the nearby people action. Choi et al. [14] propose Spatio-Temporal Volume (STV) descriptor, which captures spatial distribution of pose and motion of individuals on the scene to analyze group activity. STV descriptor centered on a person of interest is used to classify centered person's group activity. SVM with pyramid kernel is used for classification. The same descriptor is leveraged in [15], however, the random forest classification is used for group activity analysis. In [44, 14, 15], the nearby person that serves as context are selected according to the distance to the centered target and hard to guarantee the existence of interactions. To address this issue, Tran et al. [68] proposed a group-context activity descriptor similar to [44], but the people are first clustered into groups by modeling the social interaction among the individuals. However, due to the noisy observation in videos, the group detection might not be robust or stable. Therefore, our approach utilizes the social interaction region to select the contextual people without a clustering process. Besides focusing on human as context, this thesis also considers scene as a surce of contextual information, as certain activities are more likely happen at certain locations. We utilize the existing place recognition method [89] to provide scene context features that have semantic meanings.

2.7 Activity Recognition Using Deep Model

In recent years, deep models including deep neural networks, convolution neural networks, and auto-encoders have shown dramatic progress in applications like object recognition. For human activity recognition [35, 38], convolution neural networks and auto-encoder approaches [32] have been developed. However, these action/activity deep models are target-centered and do not consider any context information, which is important for human activity that involves multiple people. Comparatively, Wang et al. [74] proposed an event recognition framework, which is a hierarchical context model that captures the context information on multiple levels. Inspired by [74], our approach has the similar philosophy and uses deep structures to explicitly learn the context from people group observation and scene observation.

Chapter 3

Subject Centric Group Feature

3.1 Introduction

Person re-identification is a fundamental task in a multi-camera surveillance system to associate people across camera views at different locations and times [26]. With a growing network of cameras being used for security applications, manual re-identification that relies on a human operator is ineffective and lacking in reliability and scalability [41, 82]. Therefore, an automatic solution to person re-identification has received increasing attention from the computer vision community. Person re-identification is a challenging task and relies predominantly on visual features, such as clothing and the accessories that people carry. The visual features are intrinsically weak for matching people [26], because different people maybe dressed similarly, while the visual features of the same people may change significantly due to the changes in view angle, lighting and observed occlusions.

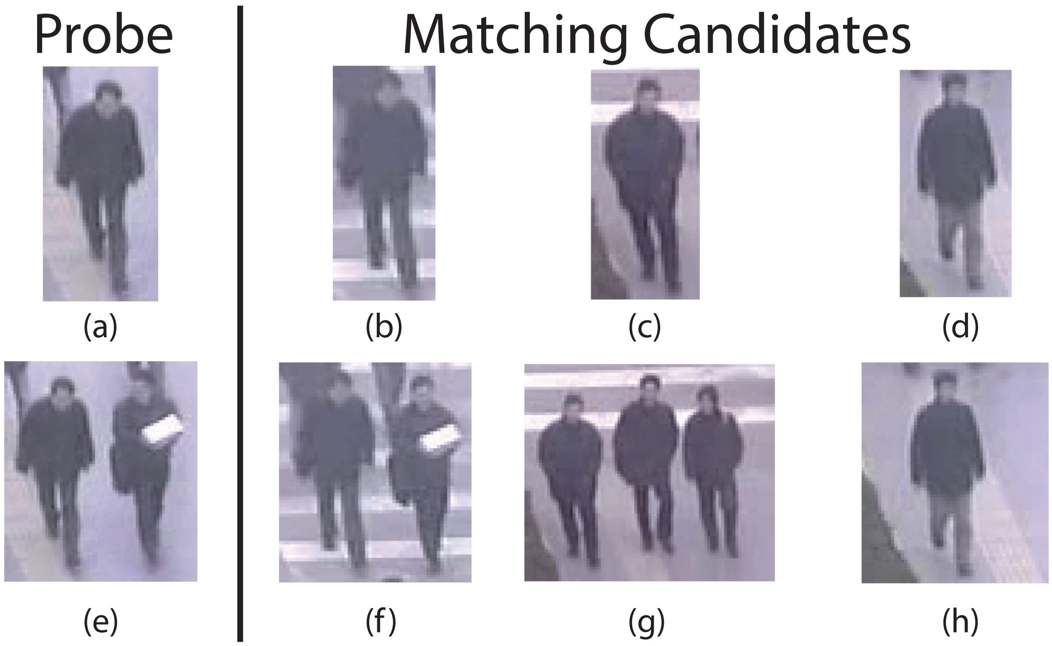


Figure 3.1: An example of group information assisting person re-identification. The first row is the persons' individual image, with (a) is the probe and (b)˜(d) are candidates that match with (a). (e)˜(h) are the group images probe person and candidates.

Many recent approaches have focused on solving the re-identification problem by developing a feature representation of a person, using low-level appearance features, such as color [84], texture [86] or their combinations [87]. Once a suitable representation is obtained, a distance metric is used to measure the similarity/dissimilarity between samples. This chapter refer to this method as the 'baseline method,' upon which it introduces the group information to improve the accuracy of re-identification.

The motivation of proposed approach is the observation that people often tend to walk alongside others or in a group. Such information can serve as context to reduce the ambiguity of person re-identification. If cameras are not geographically far apart, the same group structure could re-appear in neighboring cameras. Although the visual feature of one person could be different between cameras, by taking the co-travelers' information (geometry and visual) into consideration, it can reduce re-identification ambiguity significantly. An intuitive example is shown in Figure 3.1, where (a) is the probe and (b) to (d) are the matching candidates' images. Considering only the individual images of candidates, it is hard to point out the image that matches the same person of (a) since all persons are dressed in dark color coats and long pants. The situation would be better if we also look into persons' group context. From (e) we can observe the probe person walking with a co-traveler carrying a white object on the left side. With this information, we can tell that the first candidate has the highest possibility to match with (a). Because in (f) we can observe a person is carrying a white object walking on the left side of the first candidate, while in (g) we see that the candidate walks with two other persons and in (h) we find that the candidate walks alone.

Motivated by this example, subject centroid feature, named person-group feature, is introduced to describe the person's profile within their belonging group. By combining the person-group feature with other approaches that measure the similarity/dissimilarity between individuals, we can improve the accuracy of re-identification. The idea of matching people with group context has been explored by previous works, such as [88, 10]. The novelty of proposed feature is that it utilizes not only appearance but also the geometric attribute of groups for re-identification to improve matching accuracy. The proposed approach is unsupervised and can be applied to re-identification of subjects appearing in multiple videos. The advantage of keeping this method unsupervised is to make it simple to implement and independent to the scene of the videos. In this chapter, introduced approach is evaluated on the NLPR_MCT [1] and PRID-Group [33] datasets using videos obtained from real scenarios and find an improvement in re-identification accuracy.

The main contributions of this work include:

- A framework that can improve the baseline re-identification result using people grouping information.
- A person-group feature that encodes the person's profiles within the group, including in-group-position and co-travelers' baseline features. We also propose the metric for computing the distance between person-group features.
- A rich set of experiments to demonstrate that our approach improves the baseline results to achieve higher accuracy (around 90% matching rate at rank 5 for group members), and out-performs other re-identification methods that also utilize group information.

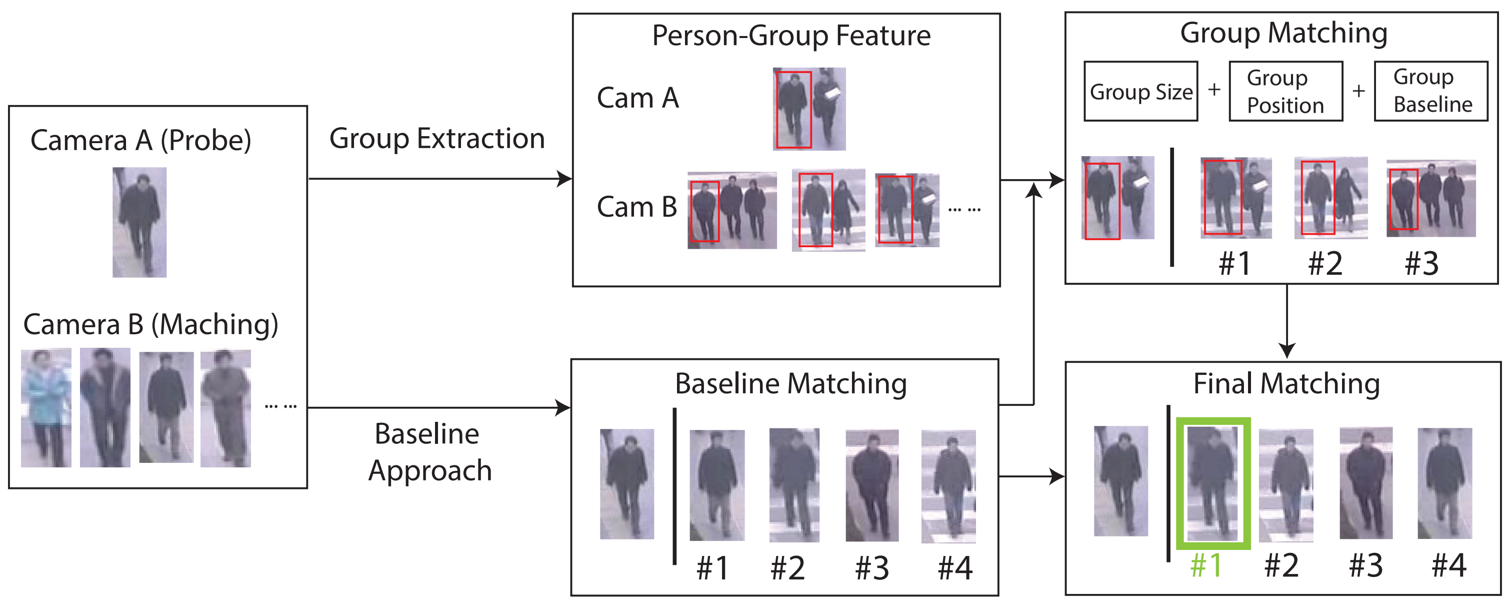


Figure 3.2: The overview of person re-identification using subject centric group feature.

3.2 Method

An overview of introduced method is illustrated in Figure 3.2. Given a probe from Camera A and a set of matching candidates from Camera B, proposed method computes a baseline matching that measures the dissimilarity score between persons from the two cameras. The baseline approach is a method that estimates the dissimilarity score of persons using the individual information only. Many features can be used in the baseline approach, such as appearance features, spatio-temporal features, and so on.

Symmetry-Driven Accumulation of Local Features (SDALF) [21] and Local Maximal Occurrence (LOMO) [50] are used as baseline approaches to calculate baseline scores in evaluation. For each person appearing in one camera, we calculate multiple baseline features of that subject across all the frames showing that person. When we compute the dissimilarity score between persons from two cameras, we simply average the baseline feature distance of all possible feature pairs between them. The baseline approach results in a pair-wise score matrix, and it serves as an initial re-identification result.

The introduced method uses group information to improve the baseline score. First, it performs group extraction (Sec. 3.2.1) to extract groups using persons' trajectories. Then the person-group features (Sec. 3.2.2) are computed for each person. Person-group feature includes the in-group-position of a subject and the spatial appearance information of group members. In group matching step, we evaluate the dissimilarity between groups by considering three aspects: group size, group position and the baseline features of group members. The final step of proposed method is combining group matching result and the baseline matching results. This approach introduces two matching result combination strategy. Person-Group Feature Distance with Group Shape (Sec. 3.2.3) consider group shape and appearance information, this strategy assumes for the case that camera is very close and group shape is unlikely to change dramatically. Person-Group Feature Distance with Group Appearance (Sec. 3.2.4) only consider group appearance information, it matches group members non-rigidly and does not assume the group shape would consist across the cameras.

3.2.1 Group Extraction

This section presents a group extraction approach by clustering the person's trajectories observed in a camera view. In proposed approach, the group is defined as a set of persons traveling together through the scene. In social science research conducted by McPhail and Wohlstein [56], they analyzed and summarized pedestrian behavior from a set of film records, and proposed the objective measure for people traveling together. The group members are determined by thresholds of difference in people's positions and velocities. Ge et al. [23] directly applied these thresholds to automatically detect small groups in crowd automatically. However, we found that directly applying threshold does not provide robust results when persons' positions and velocities are noisy because both are computed from person's on-ground trajectories, which is reconstructed from persons' tracking data. To improve the robustness of group extraction, we use a kernel function to compute the probability of person grouping over frames. Next, we use affinity propagation to discover the clusters/groups of people.

Consider the trajectory of the person P_i in the scene as a set of sequence $L_i = \{(s_i^t, v_i^t)\}$, where s_i^t and v_i^t are the person's centroid (back-projected onto the ground using estimated homography) and velocity vector of P_i at frame t. Similar to [23], we compute the aggregated pairwise grouping possibility $W = [w_{ij}]$ over-time:

$$w_{ij} = \sum_{t=0}^{\infty} \delta_{ij}^t \exp\left(-\frac{\|s_i^t - s_j^t\|^2}{2\tau_s^2} - \frac{\|v_i^t - v_j^t\|^2}{2\tau_v^2}\right) / \sum_{t=0}^{\infty} \delta_{ij}^t \tag{3.1}$$

$$\delta_{ij}^t = \begin{cases} 1 & \text{Both } P_i \text{ and } P_j \text{ appear in the sence at frame } t \\ 0 & \text{Otherwise} \end{cases}$$

where τ_s and τ_v are the thresholds of spatial and velocity difference. To identify the groups, we use clustering method to find the groups with the great internal grouping possibility. As we already compute the group probability between trajectories in Equation 3.1, we can use any clustering algorithm that takes pairwise distance/similarity as input, such as K-medoids or spectral clustering. However, both methods require the number of clusters as input, which is not easy to obtain in our problem. Therefore, we use Affinity Propagation (AP) [22] to discover both the group numbers and group members. Each person forms a data point, and the grouping possibility matrix W is used as the similarity matrix, which is the input to AP. The output of AP is a set of exemplars and corresponding clusters/groups. We denote these groups as $G = \{g_i\}$. We also use $G(P_i)$ to denote the group that P_i belongs to. Figure 3.3 shows two examples of the group extraction results.

Figure 3.3: Two examples of group extraction results. The images are video frames from two non-overlapping cameras. The persons' bounding boxes and trajectories of 2 seconds are shown on the figures. In each figure, the persons belong the same group are marketed using the same color.

3.2.2 Person-Group Feature

This section introduces the person-group feature, which describes two things about a subject within a group: who are the people that subject traveling with, and how they travel with that person. For the first part, we collect the subject's co-travelers' baseline feature and re-utilize the baseline score to evaluate the similarity of co-travelers. For the second part, we propose an in-group-position signature to encode the position of the subject within a group. We compute the local positions of co-travelers with respect to the subject's moving direction through time, the in-group-position signature is a set of co-travelers' positions. The distance measure between in-group-position signatures can be computed by solving the integer programming problem inspired by Earth Mover Distance [63].

In-group-position signature. Assume we want to construct the in-group-

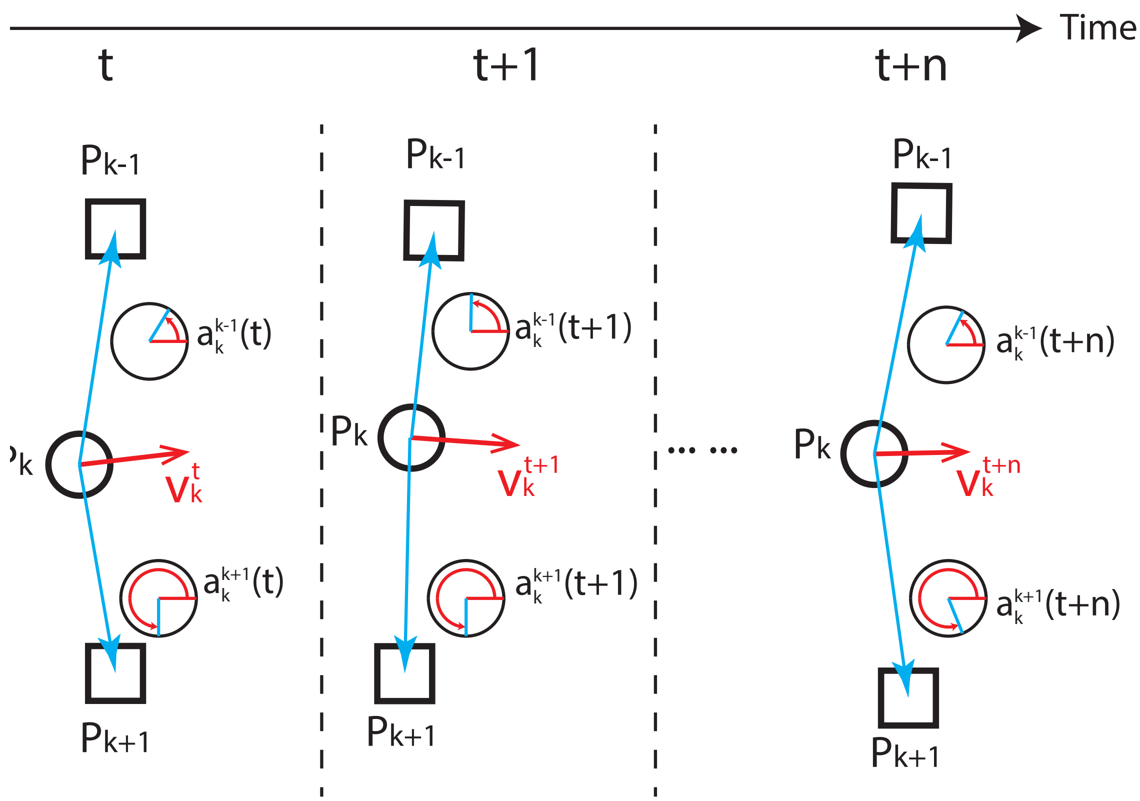


Figure 3.4: The circle denotes the centric subject; rectangle denotes the co-traveler. Red arrow points to the subject moving direction; blue arrow denotes the co-traveler direction.

position signature of P_i, where P_i belongs to group $G(P_i)$. Firstly, for each $P_j \in G(P_i)$ and $P_j \neq P_i$, we have to compute the angles between P_j and the moving direction, from perspective of P_i through all frames. We denote (s_i^t, v_i^t) as P_i's position and velocity at frame t. Then the angle between P_j and moving direction is computed as:

$$\alpha_i^j(t) = \begin{cases} \Delta & \mathbf{\Gamma} \cdot \mathbf{Z} \geq 0 \\ 2\pi - \Delta & \text{Otherwise} \end{cases} \tag{3.2}$$

$$\Delta = cos^{-1} \frac{(s_j^t - s_i^t) \cdot v_i^t}{|s_j^t - s_i^t||v_i^t|}$$

$$\mathbf{\Gamma} = v_i^t \times (s_j^t - s_i^t)$$

$$\mathbf{Z} = (0, 0, 1)$$

We collect $\alpha_i^j(t)$ through all frames, which is fitted by a Gaussian distribution, and we denote this distribution as $\alpha_i^j = (\mu_i^j, \sigma_i^j)$, where μ_i^j is the mean angle and σ_i^j is the angle deviation. An illustration of in-group-position signature is shown in Figure 3.4.

As we collect the distributions for all group members in $G(P_i)$ except centric subject P_i, it forms a distribution set that is represented as $H_i = \{\alpha_i^j | P_j \in G(P_i), P_i \neq P_j\}$, which is in-group-position signature of P_i. We denote P_j's co-travelers baseline features as $B_i = \{\beta_i^j | P_j \in G(P_i), P_i \neq P_j\}$. Hence, we represent the person-group feature of P_i as $PG_i = (H_i, B_i)$.

3.2.3 Person-Group Feature Distance with Group Shape

Given person-group features, the distance measure between features are based on a linear combination of three terms: group size distance, in-group-position distance, and group baseline distance. Let PG_i and PG_j denotes the person-group feature of P_i and P_j. Their distance takes the form:

$$D(PG_i, PG_j) = D_g(G(P_i), G(P_j)) + D_p(H_i, H_j) + D_b(B_i, B_j) \tag{3.3}$$

The first term D_g is the group size distance, which return the size difference of groups that includes P_i and P_j. The group size distance is computed by:

$$D_g(G(P_i), G(P_j)) = \| |G(P_i)| - |G(P_j)| \| \tag{3.4}$$

where $|G|$ is the group size (number of group members) of group G.

The second term D_p is the in-group-position distance, which evaluates the difference between in-group-position signatures. As we know, $H_i = \{\alpha_i^j | P_j \in G(P_i), P_i \neq P_j\}$ is a set of distributions that encode the co-traveler's location around P_i. H_i is a distribution in metric space. The problem of computing distance between H_i and H_j becomes one of computing the distance between two distributions. There are many metrics that define distance between distributions. We found that the intuition behind Earth Mover Distance (EMD) [63] fits our problem best. EMD computes the distance between distributions in space by computing minimum cost of turning one distribution to another, where costs are assumed to be amount of weights moved, times the distance by which it is moved in space. The minimum cost can be solved as a linear programming problem. In our problem, we define the distance between in-group-position signature as the minimum amount of deformations that transfer one feature to another. However, unlike the original EMD algorithm, the person can only be transformed as a complete part, therefore integer programming is required to solve the minimum deformation in our problem.

Let $H_s = \{\alpha_s^1, ..., \alpha_s^m\}$ be the in-group-position signature of P_s, $H_t = \{\alpha_t^1, ..., \alpha_t^n\}$ be the in-group-position signature of P_t. As we mentioned above, all possible angle distribution belongs to a metric space M. The distance function of M is simply defined as the distance between the distributions' mean angle:

$$Dis(\alpha_s^m, \alpha_t^n) = \begin{cases} \frac{\Theta}{\pi} & \Theta \leq \pi \\ 2 - \frac{\Theta}{\pi} & \text{Otherwise} \end{cases}$$

$$\Theta = |\mu_s^m - \mu_t^n|$$

Let $D = [d_{ij}]$ be the difference between *i-th* element in H_s and *j-th* element in H_t. We try to find a flow $F = [f_{ij}]$, where f_{ij} is a binary variable, with $f_{ij} = 1$ when *i-th* element of H_s is moved to the same location of *j-th* element in H_t after the deformation. This optimization can be formulated as a binary integer programming problem:

$$F = \arg\min_F \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \tag{3.5}$$

subjects to the following constrains:

$$
\begin{aligned}
& f_{ij} \in \{0,1\}, 0 \le i \le m, 0 \le j \le n \\
& \sum_{i=1}^{m} f_{ij} \le 1, 1 \le j \le n \\
& \sum_{j=1}^{n} f_{ij} \le 1, 1 \le i \le m \\
& \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(m, n)
\end{aligned}
$$

After we solve the above optimization, the in-group-position signature distance is calculated using:

$$D_p(H_s, H_t) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{3.6}$$

The final term D_b is the group baseline distance. It computes the aggregated differences of co-travelers' baseline features, under the condition that the co-traveler's correspondence is known by solving Equation 3.5. Let $R = [r_{ij}]$ be the pairwise baseline distance matrix, where r_{ij} denotes the baseline distance between *i-th* element in B_s and *j-th* element in B_t. The group baseline distance takes the form:

$$D_b(B_s, B_t) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} r_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{3.7}$$

When a person is traveling alone, the person-group feature is empty. In this case, the distance to an empty person-group feature D_p and D_b are set to zero and only group size distance, G_g, contributes to the person-group feature difference.

As we introduced the person-group feature and defined the distance function between features. We argue that by combining the metric of person-group feature and baseline feature, we can improve the performance of person re-identification. A simple way to combine two distance measurements is by linearly adding them:

$$D(P_i, P_j) = D(B_i, B_j) + D(PG_i, PG_j) \tag{3.8}$$

Where B_i is the baseline feature of P_i, and $D(B_i, B_j)$ means the baseline distance of person P_i and P_j.

3.2.4 Person-Group Feature Distance with Group Appearance

The approach we introduced in Sec 3.2.2 has an important assumption, which is the cameras are not far apart, and the spatial relationship among group members are preserved from one camera to another. However, in cases of cameras are far apart, this assumption hardly stands. Because there is a high probability of distraction events happen and change the group structure. Therefore, using the group shape based metric can significantly affect the performance of person-group feature in person re-identification.

To solve the limitation of unstable group structure, we introduce another matric which considers the group appearance only. The main idea is that if the same group appear in cameras views, the group members would have the similar overall appearance. The overall similarity between two sets of group members is estimated by finding a bijection that minimizes the accumulated appearance difference. This problem can be modeled as a non-rigid registration [80] using closest-appearance members between the groups.

Let PG_s and PG_t denotes the person-group feature of P_s and P_t, their group appearance based distance $M(PG_s, PG_t)$ takes the form:

$$M(PG_s, PG_t) = D_g(G(P_s), G(P_t)) + D_n(B_s, B_t) \tag{3.9}$$

Where D_g is the group size distance as defined in Equation 3.4. The second term D_n is the group appearance distance. Similar to the way we find the position matching in Equation 3.5, we also compute a binary matrix $S = [s_{ij}]$ that represent group members matching which brings the minimal aggregated appearance difference:

$$S = \arg\min_{F} \sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij} r_{ij} \tag{3.10}$$

The optimization above subjects to the following constrains:

$$\begin{aligned}
&s_{ij} \in \{0,1\}, 0 \le i \le m, 0 \le j \le n \\
&\sum_{i=1}^{m} s_{ij} \le 1, 1 \le j \le n \\
&\sum_{j=1}^{n} s_{ij} \le 1, 1 \le i \le m \\
&\sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij} = \min(m, n)
\end{aligned}$$

where r_{ij} denotes the baseline distance between *i-th* element in B_s and *j-th* element in B_t. After we solve the optimization, the group appearance distance is calculated using:

$$D_n(B_s, B_t) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij} r_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij}} \tag{3.11}$$

When we calculate the distance between two persons using group appearance-based approach, similar to group shape-based matching, we combine two distance measurements is by linearly adding them:

$$D(P_i, P_j) = D(B_i, B_j) + M(PG_i, PG_j) \tag{3.12}$$

Where B_i is the baseline feature of P_i, and $D(B_i, B_j)$ means the baseline distance of person P_i and P_j.

3.3 Experiments and Comparisons

3.3.1 Evaluation Datasets

To evaluate our approach, we test our method on the NLPR_MCT [1] and PRID-Group [33] dataset. Other re-identification datasets (e.g. CAVIAR, VIPeR, ETHZ) either contain single person's images only or do not have group information provided in the dataset. Therefore they are not suitable for evaluation of our approach. The Dataset 1 and 2 of NLPR_MCT is used for evaluation. For both datasets, there are three synchronous videos (resolution: 320×240, 20 frames-per-second) from three non-overlapping cameras. We use the videos produced by two outdoor cameras for evaluation. The number of people in each dataset is presented in Table 3.1. The dataset provides the ground-truth annotation, which includes the bounding box tracking for each people. To better evaluate our approach, we create PRID-Group as an additional dataset to evaluate our subject

Figure 3.5: Two Examples of Re-identification Results. For each query, the image of query person and the group that person belongs to are presented. We display the matching results of baseline approach [21] and our approach. We display the top four candidates in the table, as well as the group each candidate belongs to in the same grid. The ground truth matching is labeled by blue boxes, where the rank is also given at right. The ranks with star symbols are the results obtain using proposed approach. Otherwise, the ranks are computed by baseline approach.

	Camera 1	Camera 2	Common
NLPR_MCT Dataset 1	76	78	72
NLPR_MCT Dataset 2	115	111	105
PRID-Group	38	38	38

Table 3.1: People Number of Evaluation Datasets

centric group feature. PRID dataset [33] contains two synchronous videos (resolution: 720×576, 25 frames-per-second) of two non-overlapping street views. It provides tracking information of 200 individual from two videos for re-identification task evaluation. In the original dataset, most of individuals walk alone. A small subset of individuals walk in groups, but most group members are not recorded by the dataset due to occlusion among co-walkers. Therefore, the original dataset is not suitable for group based re-identification evaluation. In order to record the persons traveling in groups, we find groups of person by observing the persons location and interaction in video, then manually annotate 38 individuals, which forms 16 groups appearing in both camera views. The tracking information of each person is computed using [58]. The challenge of person re-identification of PRID-Group dataset comes in three folds: First, the videos from two cameras contains a large viewpoint change; second, there is a stark difference in illumination, background and camera characteristics between the two videos; third, occlusion among group members is frequent in the first video. Some example group images are shown in Figure 3.11.

In both NLPR_MCT and PRID-Group dataset, the persons' X-Y plane locations are computed by back projecting the mid-bottom of bounding boxes, and the homography is estimated interactively, off-line. The group information of

NLPR_MCT dataset is extracted using proposed algorithm in Section 3.2.1. In dataset 1, both Camera 1 and Camera 2 have 18 persons traveling with co-travelers and form 8 groups (with size greater than 1). In dataset 2, 35 and 31 persons walk with co-travelers, and they form 16 and 15 groups in Camera 1 and Camera 2, respectively.

We measure the performance using Cumulative Match Curve (CMC) [28], we calculate the person-group feature distance from persons appeared in one camera to another, and sort the persons in ascending order based on the distance value. The rank score is the order of ground-truth match in the sorted person list. We demonstrate some examples of query and candidates person/group images in Figure 3.5. The results obtained using [21] are also provided.

3.3.2 Features Evaluation

In this section, we evaluate how group extraction and each component of subject centric group feature contributes to the final accuracy of the re-identification performance.

Group Extraction Evaluation. Since our approach depends on the group information given by the group extraction method, we want to discover how different group extraction algorithms affect the re-identification results. We choose Ge et al. [23] as the comparing group extraction method. The results is illustrated in Figure 3.6. As seen, using group information extracted by either method lead to an improvement in accuracy compared to the baseline approach. In Dataset 1, our

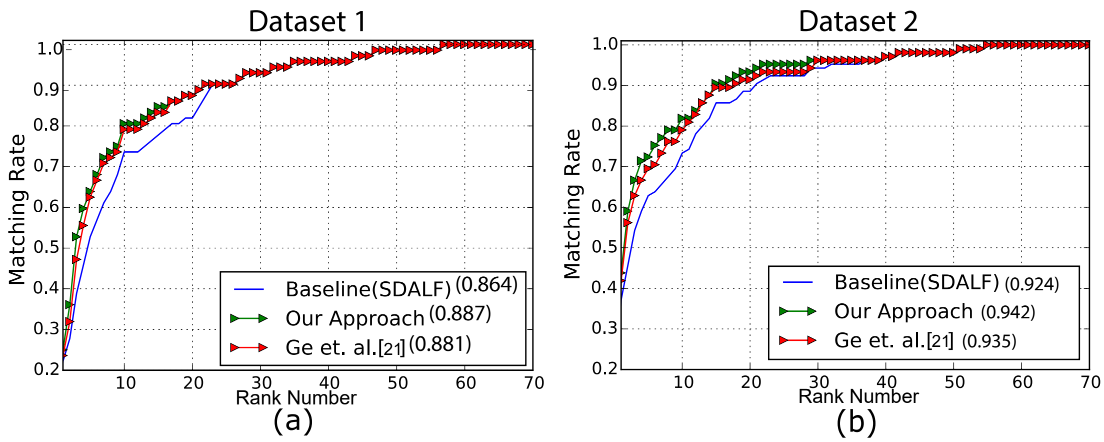


Figure 3.6: CMC curves for person re-identification with group information extracted using our approach and that of Ge. et al. [23]. The value of the normalized area under the CMC curve (uAUC) is given in the parentheses.

approach provides similar accuracy as with [23], while in Dataset 2, our approach is slightly better. The reason is that Dataset 1 has less crowded scenes and the group extraction task is relatively easier. However, the scenes are more crowded in Dataset 2 and the performance of [23] is effected by directly using the threshold and noisy trajectories. Our group extraction algorithm handles noise better by computing the grouping probability using a kernel function, which leads to more accurate group information and benefits the re-identification task.

Group Features Evaluation. In group shape based metric, as it shows in Equation 3.3, the distance between person-group features contains three terms: group size (GS), in-group-position (GP), group baseline (GB). In group appearance based metric, as it shows in Equation 3.9, the distance contains two terms: group size (GS) and group appearance (GA).

We evaluate the contribution of each term by showing how well the person

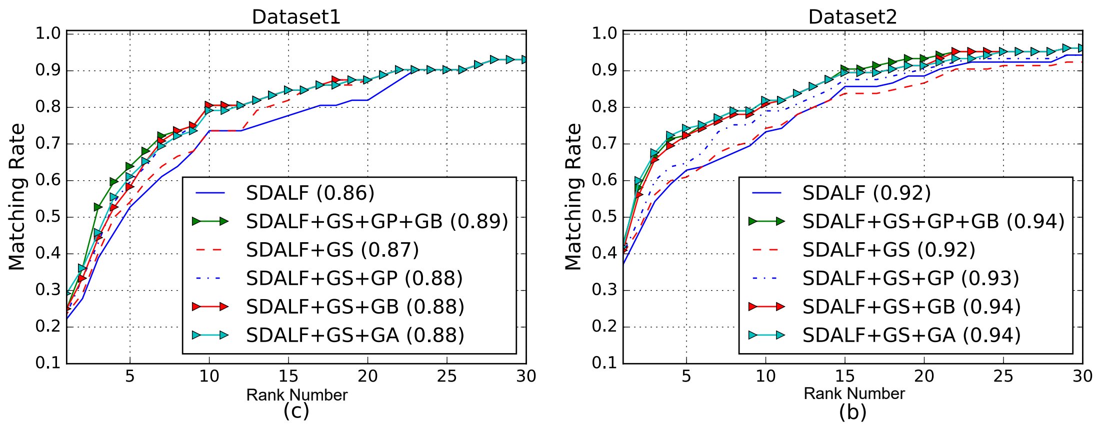


Figure 3.7: The CMC evaluation of group size(GS), in-group-position(GP) and group baseline(GB) of person group feature.

re-identification performs when each part is included in the final distance. We use SDALF as the baseline and show the CMC curve of GS, $GS+GP$, $GS+GB$, $GS+GA$(Group appearance-based metric) and $GS+GP+GB$(Group shape-based matric). The reason that we combine GP, GB and GA with GS is that GP, GB, and GA are only meaningful to compare the group with the same size, and we have to combine them with GS to calculate the distance between groups of different size. The results are illustrated in Figure 3.7, and the rate comparison is given in Table 3.2. As seen in both Dataset 1 and 2, the GS distance alone provides very limited improvement. This is because GS only consider the size of the group and does not take any appearance information into account. We also observe that both $GS+GP$ and $GS+GB$ perform better than GS as they consider the location and appearance information of the group members. The performance of $GS+GB$ is slightly better than $GS+GP$, the reason being that GB is calculated based on the in-group-position mapping of GP and it leads to less ambiguity by considering the person's appearance in GB.

We also observe that the overall performance of group-appearance based metric $GS+GA$ is very close to group-shape based metric $GS+GP+GB$. In Dataset 1, $GS+GA$ performs slightly better than $GS+GP+GB$ at low rank, however, $GS+GA$ becomes slightly worse at higher ranks. In Dataset 2, their performance are similar across all ranks. The reason being that group persons do not change their position significantly across the camera views in the testing dataset. This fact allows both approaches finding similar group matching when calculate the person group feature distance, and leads to a very similar performance.

Dataset	Rank	SDALF	SDALF+ GS	SDALF+ GS+GP	SDALF+ GS+GB	SDALF+ GS+GA	SDALF+ w/ GS+GP+GB
1	1	0.22	0.24	0.24	0.25	0.29	0.25
	5	0.52	0.54	0.60	0.58	0.61	0.64
	10	0.74	0.74	0.80	0.80	0.79	0.80
2	1	0.37	0.40	0.40	0.41	0.42	0.43
	5	0.62	0.61	0.64	0.72	0.74	0.72
	10	0.73	0.74	0.79	0.81	0.82	0.82

Table 3.2: The matching rates comparison of SDALF baseline score combined with group size(GS), in-group-position(GP), group baseline(GB) and group appearance(GA) of person group feature.

3.3.3 Compare with Baseline Approaches

To test the performance of our method under difference baseline methods, we conduct experiments using the Symmetry-Driven Accumulation of Local Features (SDALF) [21] and Local Maximal Occurrence (LOMO) [50]. SDALF requires background subtraction, which is obtained using ViBe [5]. The results are shown in Figure 3.8.

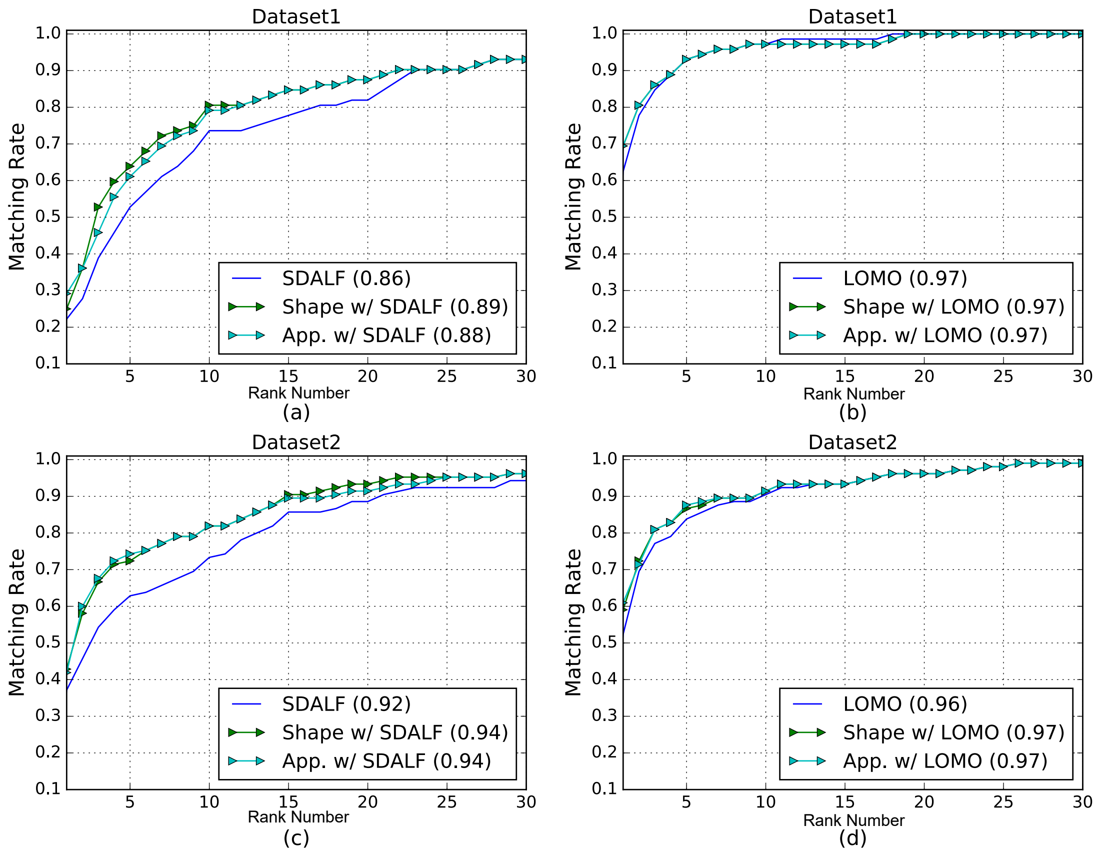


Figure 3.8: The comparison of CMC using baseline methods SDALF and LOMO on two datasets. The value of uAUC is shown with in the parentheses.

Dataset	Rank	SDALF	Shape w/ SDALF	App. w/ SDALF	LOMO	Shape w/ LOMO	App. w/ LOMO
1	1	0.22	0.25	0.29	0.62	0.69	0.69
	5	0.52	0.64	0.61	0.93	0.93	0.93
	10	0.74	0.80	0.79	0.97	0.97	0.97
2	1	0.37	0.43	0.42	0.52	0.59	0.61
	5	0.62	0.72	0.74	0.83	0.87	0.88
	10	0.73	0.82	0.82	0.90	0.91	0.91

Table 3.3: The matching rates comparison between our approach and baseline methods (SDALF and LOMO)

The matching-rates comparison among our group shape-based, appearance-based approaches and baseline methods at rank 1, 5, and 10 are given in Table 3.3. Both our group shape-based and appearance-based approaches outperform the baseline method, while the improvements are more significant in Dataset 2 than Dataset 1. This can be attributed to the fact that there are more persons traveling with co-travelers in Dataset 2. We also observe that the improvement of SDALF is more significant than LOMO. The reason is that our approach mainly boosts the re-identification accuracy of those who travel within groups. The accuracy improvement is bounded by the maximal possible accuracy improvement of group members. The re-identification of group members IDs using SDALF as baseline feature are shown in Figure 3.9 (b) and (f), the accuracy of each approach at rank 1, 5 and 10 are given in Table 3.4. The baseline approach does not perform well on group members IDs and leave room for improvement by using group information. Using our group shape-based approach, the improvements are 11%, 43% and 27% at rank 1, 5 and 10, respectively, in Dataset 1, and 20%, 33% and 29% at rank 1, 5 and 10, respectively, in Dataset 2; using our group appearance-based approach, the improvements are 28%, 32% and 22% at rank 1, 5 and 10, respectively, in Dataset 1, and 16%, 39% and 30% at rank 1, 5 and 10, respectively, in Dataset 2. The re-identification of group members IDs using LOMO as baseline feature are shown in Figure 3.9 (d) and (h). As seen, LOMO performs much better than SDALF and already reaches high rank-one accuracy. By introducing group information, it is not able to gain as much improvement as SDALF, especially at high ranks. By using LOMO as the baseline, our group shape-based approach has the improvements of 22%, 0% and 0% at rank 1, 5 and 10, respectively, in Dataset 1, and 23%, 10% and 3% at rank 1, 5 and 10, respectively, in Dataset 2. Our group appearance-based approach has the improvements of 22%, 0% and 0% at rank 1, 5 and 10, respectively, in Dataset 1, and 29%, 13% and 3% at rank 1, 5 and 10, respectively, in Dataset 2.

3.3.4 Compare with Group based Approaches

We also compare our approach to [88] and [10], both of which use group information as context to improve the accuracy of individual re-id. The first approach [88] extracts Center Rectangular Ring Ratio-Occurrence (CRRRO) descriptor as group context feature from a manually selected static group image. Although our dataset consists of videos, we generate group images by cropping the video frames that includes all group members and computing CRRRO representations of the groups across multiple frames. When we compute the CRRRO score of two people, we use the average the score of all possible CRRRO feature pairs between the groups that the two individuals belong to. The distance between CRRRO features is linearly combined with other appearance-based distance as the final score. The second work uses Relative Appearance Context (RAC) feature as group context, which measures the appearance difference of person to the nearby people. The distance of appearance feature is also linearly combined with relative appearance context distance as the final distance value. To make sure the comparison is fair, in both comparison methods we use both SDALF and LOMO to represent the individual appearance feature. We use parameters as suggested by respective authors

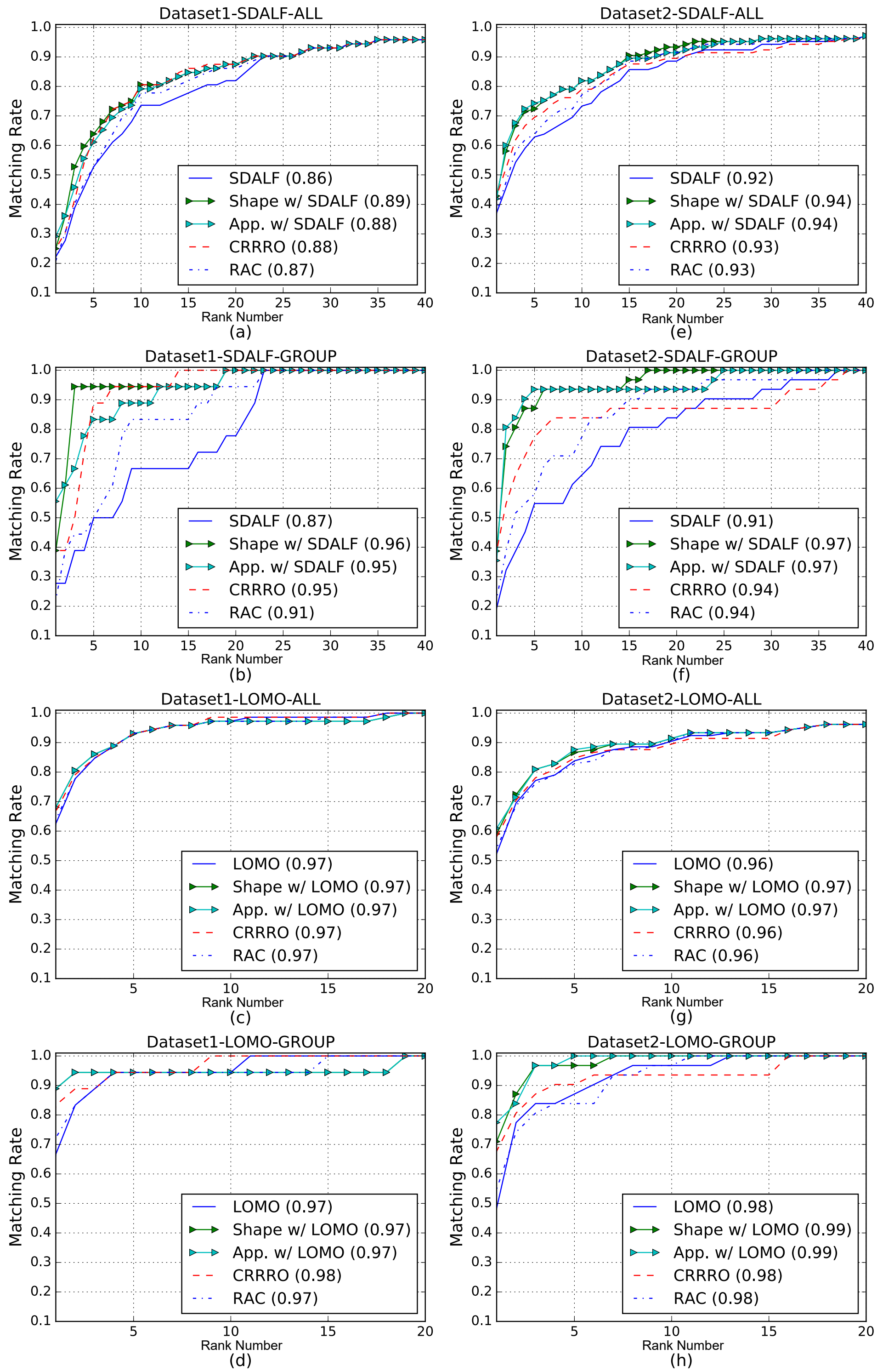


Figure 3.9: Compare the CMC of person re-identification using our approach, CRRRO descriptor and RAC feature.

in all our experiments.

Results are as shown in Figure 3.9. The matching rates at rank 1, 5, and 10 are given in Table 3.4. As seen, there is an overall improvement in re-identification accuracy. To further evaluate the impact of the person-group feature, we specifically restrict the dataset to those ID's that are found in a group. Results obtained on this restricted dataset are as shown in Figures 3.9(b) and (f) for SDALF baseline, and Figures 3.9(d) and (h) for LOMO baseline. As can be seen from the results, our method provides the best performance in both datasets. By looking into the CMC for all persons (Figure 3.9(a), (c), (e) and (g)), we can observe that the accuracy is boosted through our approach. In general, the accuracy is slightly better than compared approaches. However, as seen through the CMC of group persons (Figure 3.9(b), (d), (f) and (h)), our method can reach the accuracy of around 90% at rank 5 using SDALF baseline, which is significantly better than the baseline method and compared approaches. It can reach around 95% using LOMO baseline, which is also better than compared approaches.

To further evaluate proposed approach, we conduct re-identification on PRID-Group dataset, which contains only the people traveling within groups, both SDALF and LOMO are used as baseline features. Example probes and top four matching results from the PRID-Group data with SDALF baseline are shown in Figure 3.11. The matching rate at Rank 1, 5 and 10 are shown in Table 3.4. The result shows that SDALF reaches 39% matching rate at rank 10. However,

		Baseline = SDALF					Baseline = LOMO				
DataSet	Rank	Baseline	Shape w/ Baseline	App. w/ Baseline	CRRRO	RAC	Baseline	Shape w/ Baseline	App. w/ Baseline	CRRRO	RAC
1-ALL	1	0.22	0.25	0.29	0.25	0.21	0.62	0.69	0.69	0.67	0.64
	5	0.52	0.64	0.61	0.62	0.53	0.93	0.93	0.93	0.93	0.93
	10	0.74	0.80	0.79	0.80	0.78	0.97	0.97	0.97	0.97	0.97
1-Group	1	0.28	0.39	0.56	0.38	0.22	0.67	0.89	0.89	0.83	0.72
	5	0.51	0.94	0.83	0.88	0.51	0.94	0.94	0.94	0.94	0.94
	10	0.67	0.94	0.89	0.94	0.83	0.94	0.94	0.94	1.00	0.94
2-ALL	1	0.37	0.43	0.42	0.42	0.38	0.52	0.59	0.61	0.58	0.54
	5	0.62	0.72	0.74	0.70	0.63	0.83	0.87	0.88	0.82	0.83
	10	0.73	0.82	0.82	0.79	0.77	0.90	0.91	0.91	0.90	0.90
2-Group	1	0.19	0.39	0.35	0.39	0.23	0.48	0.71	0.77	0.68	0.55
	5	0.54	0.87	0.93	0.77	0.58	0.87	0.97	1.00	0.90	0.84
	10	0.64	0.93	0.94	0.84	0.77	0.97	1.00	1.00	0.94	0.97
PRID-Group	1	0.11	0.16	0.16	0.16	0.13	0.38	0.59	0.59	0.57	0.46
	5	0.32	0.74	0.71	0.66	0.45	0.49	0.95	1.00	0.81	0.76
	10	0.39	0.87	0.92	0.76	0.63	0.78	1.00	1.00	0.89	0.86

Table 3.4: Comparison of NLPR_MCT and PRID-Group dataset matching rates across methods that use group information for re-id.

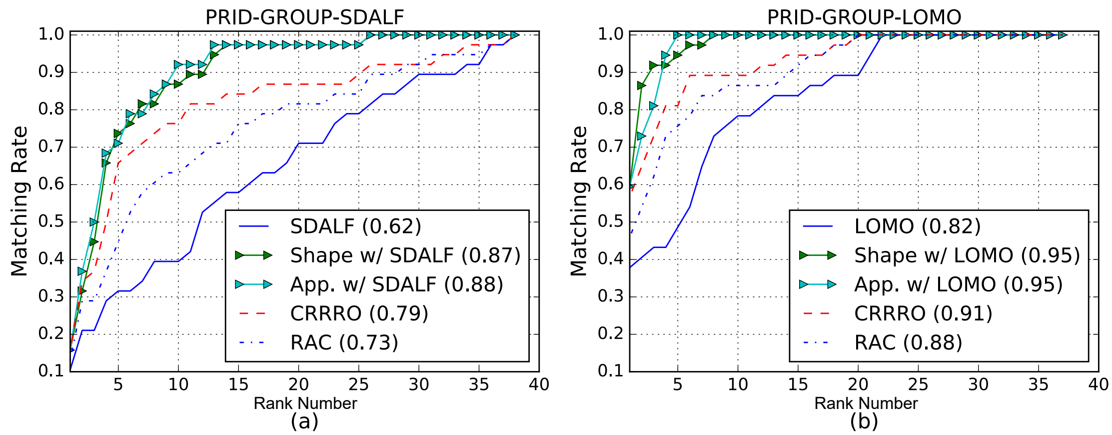


Figure 3.10: Compare the CMC of PRID-Group person re-identification using our approach and comparing approaches.

by introducing group information as additional information, all group based re-identification approaches show an improvement in matching accuracy in this experiment. In this experiment, RAC reaches 13%, 45% and 63% at rank 1, 5, and 10, respectively. CRRRO reaches 16%, 66% and 76% at rank 1, 5 and 10, respectively. Our group-shape based approach has the better performance compare to CRRRO and RAC, and reaches 16%, 74% and 87% at rank 1, 5 and 10, respectively; Our group-appearance based approach also able to reaches 16%, 71% and 92% at rank 1, 5 and 10, respectively. LOMO baseline approach does much better than SDALF baseline, and reaches 38%, 49% and 78% at rank 1, 5, and 10, respectively, and our group-shape based approach has the better performance and reaches 59%, 95% and 100% at rank 1, 5 and 10, respectively; Our group-appearance has even the better performance compare to others and reaches 59%, 100% and 100% at rank 1, 5 and 10, respectively.

Figure 3.11: Two Examples of Re-identification Results with SDALF baseline in PRID-Group dataset. The ground truth matching is labeled by blue boxes, where the rank is also given at right. The ranks with star symbols are the results obtain using our approach. Otherwise the ranks are computed by baseline approach.

3.4 Discussion

In this chapter, we address the problem of person re-identification using subject centric group features. We proposed person-group feature that encodes the geometry and visual information of groups. The distance between person-group features is computed by solving an integer programming problem. The final distance is a linear combination of person-group feature distance and a baseline distance obtained by considering appearance features. We demonstrate that our proposed method can always improve the accuracy of a baseline approach, and outperform the state-of-the-art group information based re-identification approaches.

One limitation is that the method assumes a reasonable crowd density, where robust group extraction algorithm can succeed.

Another limitation of proposed algorithm is that it assumes the stable group structure across the camera views. This is a valid assumption when the camera views are close to each other, but limits the application of this method on the camera networks that contains cameras far apart. Next chapter introduce present spatial appearance group feature that trains a machine learning model that is able to capture both group appearance and structure change across the camera.

Chapter 4

Spatial Appearance Group Feature

4.1 Introduction

Person re-identification is the task that associates humans that are observed across multiple-cameras at different locations and times [26]. As a growing network of the camera are being deployed for security applications, automatic person re-identification has received increasing attention in the computer vision community since the traditional human operator is ineffective and lacking in reliability and scalability [41, 82].

Person re-identification in videos is very challenging. Human operators and many automatic re-identification algorithms rely predominantly on visual features, such as color [84], texture [86] or their combinations [87]. However, the visual cues are weak for matching people because the difference of viewpoint and lighting condition among location and times could significantly affect the visual appearance of the same person. Other than that, extracting meaningful visual features of the individual from security video can be difficult because the wide range of video view angles and occlusions. Finally, the appearance feature also suffers from inter-similarity as different people may be wearing similar clothes.

This chapter is targeting the re-identification task between two cameras. We assume that two cameras are not far away and that the persons who travel together in one camera remain as a group in another camera. The group appearance and relatively location among group members can serve as additional evidence to reduce the ambiguity of re-identification task. More specificity, We introduce spatial appearance group (SAG) feature, which captures both the group appearance and groups structure around a subject. We learn the appearance model to predict the likelihood of visual features belonging to the same person; we also train a group model to estimate the probability two spatial appearance group feature sets describing the same group.

We demonstrate that our approach outperforms other group information based approaches. In the following of this chapter, we discuss details of the method in Section 4.2, then we show the experiments results and comparison at Section 4.3.

4.2 Method

The overview of introduced approach is shown in Figure 4.1. This method starts with a training dataset that contains the videos clips of a set of persons that

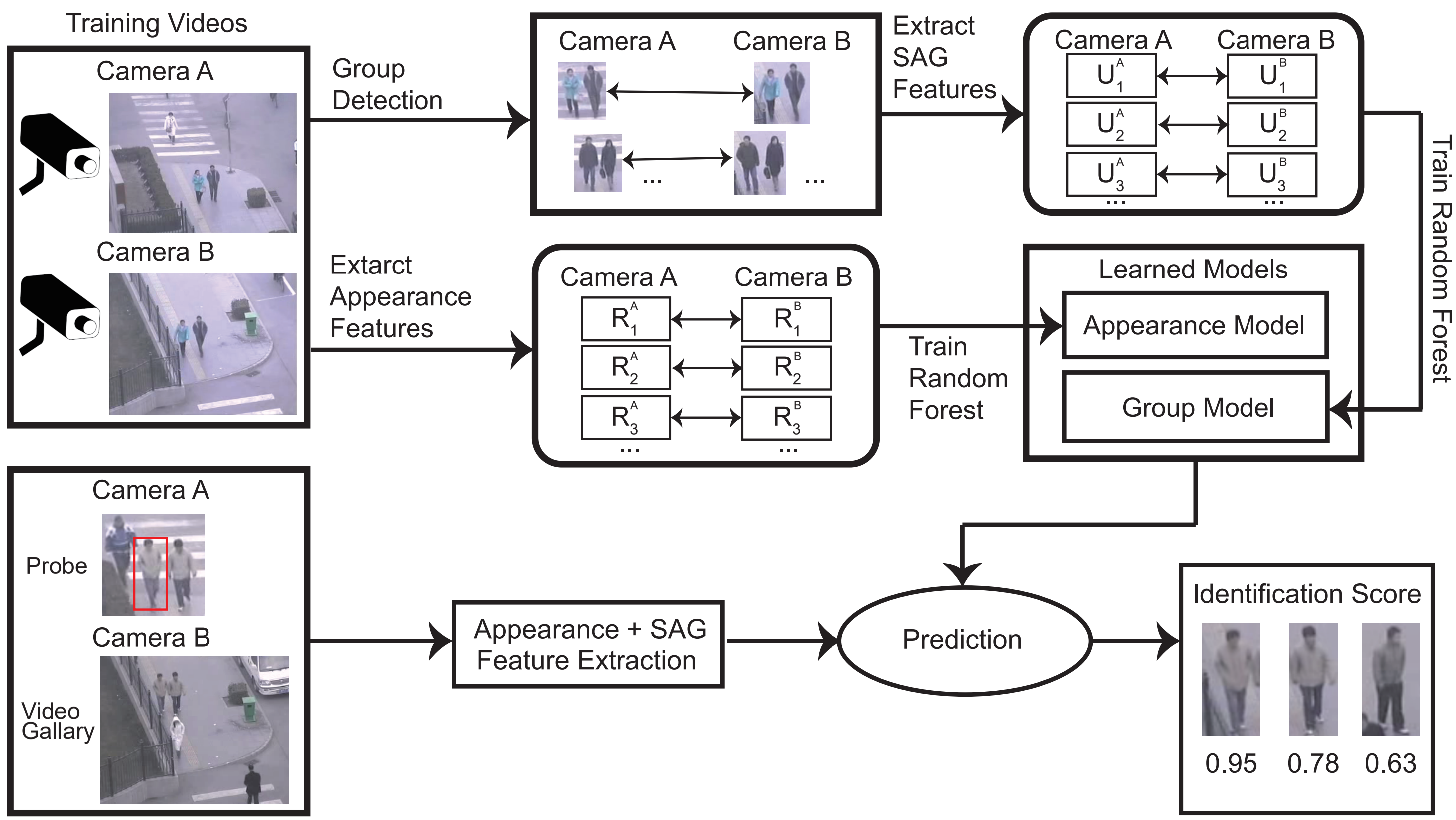


Figure 4.1: The Overview of Person Re-identification using Spatial Appearance Group Feature.

captured by two cameras. We assume that the person tracking information is available in the Dataset. We denote the tracking information of the i^{th} person in camera A and camera B as P_i^A and P_i^B, respectively. Given the training data, we extract appearance features as well as spatial appearance group features to train two models: appearance model and group model. The appearance model uses a random-forest classifier to predicts the probability of two given appearance feature sets are extracted from the same person. The group model uses another random-forest classifier that predicts how likely is it that the two observed people are the same person given two sets of spatial appearance group features. After we train the models, we can predict the re-identification score of given probes and galleries by linearly combining the appearance score and group.

4.2.1 Appearance Model

Appearance model estimates the probability of two appearance feature sets coming from the same person observed by two cameras. We seek a function H_a that:

$$H_a : (R_i, R_j) \rightarrow Y \tag{4.1}$$

The appearance feature set $R_i = \{r_i^t\}$, where r_i^t is a M dimension appearance feature vector of a person P_i at time t. In this paper, we utilize SDALF [21] to extract the appearance feature of person. Function H_a returns a scale value $Y \in [0, 1]$. We want to find H_a that returns $Y = 1$ when the R_i and R_j are the appearance feature sets from the same person, and returns $Y = 0$ otherwise. We define H_a as:

$$H_a(R_i, R_j) = \frac{1}{|R_i||R_j|} \sum_{\substack{r_i^t \in R_i \\ r_j^t \in R_j}} h_a(r_i^t, r_j^t) \tag{4.2}$$

where $h_a : (r_i, r_j) \to \{0, 1\}$. What we need to train is the binary classification model h_a that predicts 1 when r_i and r_j come from the same person, and predicts 0 otherwise.

Appearance Model Training. The classification model h_a is trained by supervised learning. We create a dataset using the videos from Cameras A and Camera B, in which a set of persons are captured by both cameras. Our dataset contains positive data samples and negative data samples. The positive samples, D^+, includes all possible appearance features extracted from the same person in Camera A and B. We denote positive samples as

$$D_a^+ = \{[r_i^A, r_i^B] | r_i^A \in R_i^A, r_i^B \in R_i^B\} \tag{4.3}$$

The negative samples, D^-, are the appearance features pairs that comes from different persons in camera A and camera B.

$$D_a^- = \{[r_i^A, r_k^B] | r_i^A \in R_i^A, r_k^B \in R_k^B, i \neq k\} \tag{4.4}$$

As the amount of D_a^- may the significantly greater than D_a^+ if we include all possible combinations, the resultant set can be an unbalanced dataset. We randomly pick a subset of all possible negative pairs and make $|D_a^-| = 2|D_a^+|$.

We seek the model h_a by training a random-forest classifier given D^+ and D^-. In current implementation, the positive and negative samples are weighted, and the forest contains at most 20 trees with maximal depth equal to 10.

4.2.2 Group Model

The group model predicts the probability of how likely the observed group that a subject belongs to is the same group. In this sections, we introduce the spatial appearance group (SAG) feature that encodes the group information, which includes group appearance and group structure. We also train a random-forest classifier to predict whether a pair of SAG features comes from the same group. We further utilize this classifier to compute the probability of two group features being extracted from the same group.

Group extraction. To extract group feature, we need to know the persons that belong to a group first. Assuming the tracking information for each person is available, we use the group extraction method proposed in Wei et al.[77] to detect groups. It first computes the group probabilities of all possible person pairs in according to the affinity of persons' trajectories. Then affinity propagation clustering [22] is used to find the person clusters. Each cluster is treated as one group. In our approach, we denote i^{th} group in camera A as G_i^A and denote $P_i^A \in G_i^A$ if the i^{th} person in camera A belongs to the i^{th} group.

Spatial appearance group feature. The spatial appearance group feature (SAG) feature is centered on the subject, and it captures the group appearance to describe the group structure using the spatial relationship between the focal subject and other group members.

For person P_i belonging to group G, we denotes its co-travels within the same groups as $N_i = \{P_k | P_k \in G, k \neq i\}$. We divide the context region of P_i into K sub-polar context regions characterized by a number of orientation bins. Using the spatial relationship between P_i and N_i, we compute the feature for sub-polar context region $C_i(k)$ by accumulating the appearance feature of $P_k \in N_i$ within the context area.

As P_i may appear in multiple frames in the video, we denotes P_i's group feature as $U_i = \{u_i^t\}$. u_i^t is a $K \times M$ dimension vector that represents spatial appearance feature at time t, which we compute as follows:

$$u_i^t = [\sum_{P_j \in N_i} w_{1j}^t a_j^t, \sum_{P_j \in N_i} w_2^t a_j^t, , ..., \sum_{P_j \in N_i} w_K^t a_j^t,] \tag{4.5}$$

where the weight term w_{kj}^t denotes the contribution by P_i to region $S_i(k)$ at time t, as it is shown in Figure 4.2. We compute $w_{kj}^t = |b_j^t \cap C_i(k)|/|b_j^t|$, in which b_j^t is the bounding box of P_j at time t.

Group Model Training. We need to create data samples to train the binary classification model h_g. Similar to the appearance model training, our training sample contains positive samples D_g^+ and negative samples D_g^-. We obtain positive samples D_g^+ by pairing the SAG features of the same group members that appear in both camera A and B.

$$D_g^+ = \{[u_i^A, u_i^B] | u_i^A \in U_i^A, u_i^B \in U_i^B, P_i \in \forall G\} \tag{4.6}$$

We obtain the negative samples D_g^- by pairing the SAG features of different group members that appear in both camera A and B.

$$D_g^- = \{[u_i^A, u_j^B] | u_i^A \in U_i^A, u_j^B \in U_j^B, P_i, P_j \in \forall G, i \neq j\} \tag{4.7}$$

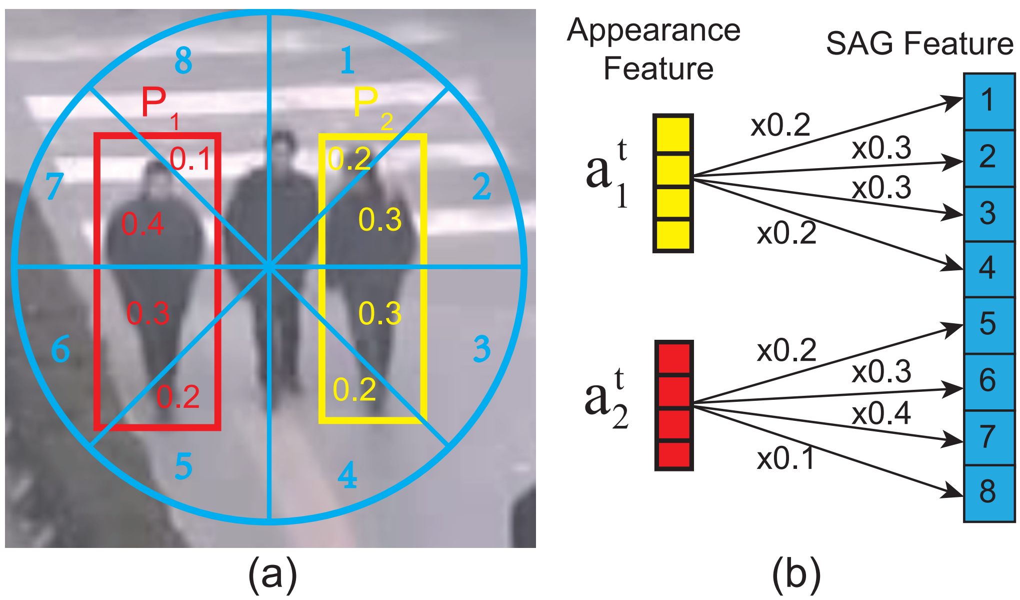


Figure 4.2: The calculation of Spatial Appearance Group Feature. (a) The blue circle denotes the context regions of the centric subject, the persons bounded by red and yellow boxes are co-traveler of the centric subject. The values inside bounding box indicate the weights of appearance feature that contributes to sub-polar context regions. (b) The SAG feature is computed by adding the weighted appearance feature of each co-traveler.

To balance the number of positive and negative samples, we generate additional positive samples by calculating additional SAG features with slightly rotated (-5 and 5 degrees) context regions of the group members, as well as sub-sample the negative samples. We make the number of negative samples to be twice as many as positive samples $|D_g^-| = 2|D_g^+|$.

Finally, we train a random-forest classifier h_g. The positive and negative samples are weighted, and we set the maximal number of classification trees in the forest to 20 and maximal tree depth to be 10 in our implementation.

4.2.3 Score Generation

After both appearance model and group model are trained, we compute the score S that indicates how similar is a pair of features extracted from persons observed in different cameras. Given person P_i and P_j, the score S_{ij} is computed as:

$$S_{ij} = H_a(R_i, R_j) + w_g H_g(U_i, U_j) \tag{4.8}$$

Where R_i and R_j are the appearance feature sets extracted and U_i and U_j are the group feature. We use w_g to adjust the weight of group score. In our experiments we simple use $w_g = 1$.

4.3 Experiments and Comparisons

This section represents the experiments and results to demonstrate the performance of proposed approach. We also compare our method with other state-of-the-art group based re-identification methods.

4.3.1 Evaluation Dataset

We evaluate our approach using dataset NLPR_MCT [1]. Other re-identification datasets (e.g., CAVIAR, VIPeR, ETHZ), they either contain single person's image or do not have reasonable numbers of groups appearing in a scene. We do our experiments on dataset 1 and 2 of NLPR_MCT where each dataset contains three synchronous videos (resolution: 320×240, 20 frames-per-second) from three non-overlapping cameras. Only two cameras that record the outdoor scenes are used in our experiments. The dataset provides annotation information and bounding box trackers for each person appearing in the video. We apply the group detection algorithm as described in Section 4.2.2 to obtain the groups in the dataset. An example of the NLPR_MCT dataset that showing a group of two persons traveling across two cameras is shown in Figure 4.3.

NLPR_MCT Dataset 1 contains 78 persons, in which 16 group members forms eight groups. NLPR_MCT Dataset 2 contains 104 persons, in which 24 group members forms ten groups. For each dataset, we split the persons into 4-folds, each fold contains equal numbers of group members and non-group members. We

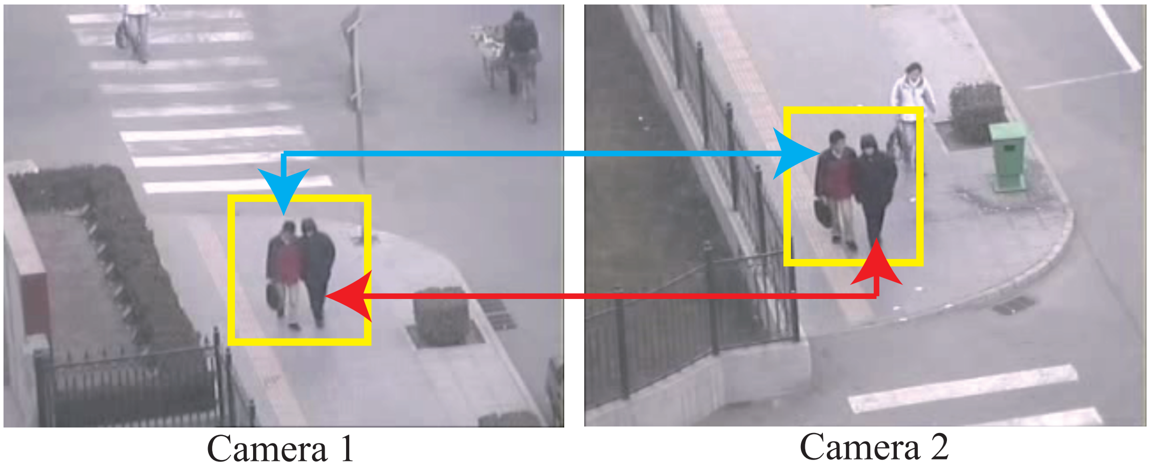


Figure 4.3: An example of two persons traveling across cameras 1 and 2. The yellow box indicates that two people form a group, blue and red arrows indicate the same person across two cameras.

Dataset(Accuracy%)	Rank	Appearance+ +SAG+RF	Appearance +RF	Appearance SAG+Euc	Appearance +Euc	CRRRO[23]	SCG[24]
NLPR_MCT dataset 1	1	55.6	41.6	52.7	45.8	40.2	50.0
	5	88.9	81.9	86.1	83.3	86.1	86.1
	10	97.2	98.6	98.6	95.8	94.4	93.1
NLPR_MCT dataset 2	1	64.4	51.9	46.1	39.4	55.7	65.3
	5	96.1	89.4	81.7	73.0	86.5	88.4
	10	97.1	98.1	90.3	84.6	97.1	93.2

Table 4.1: The matching rates (at rank 1, 5, and 10) of our approach compare to others.

use 4-folds cross-validation to report the performance where we choose one fold as testing data and use the rest as training data. The testing results are represented by Cumulative Matching Characteristic (CMC) curve. The averaged CMC curves of all folds are used to represent the overall performance of methods on the entire dataset.

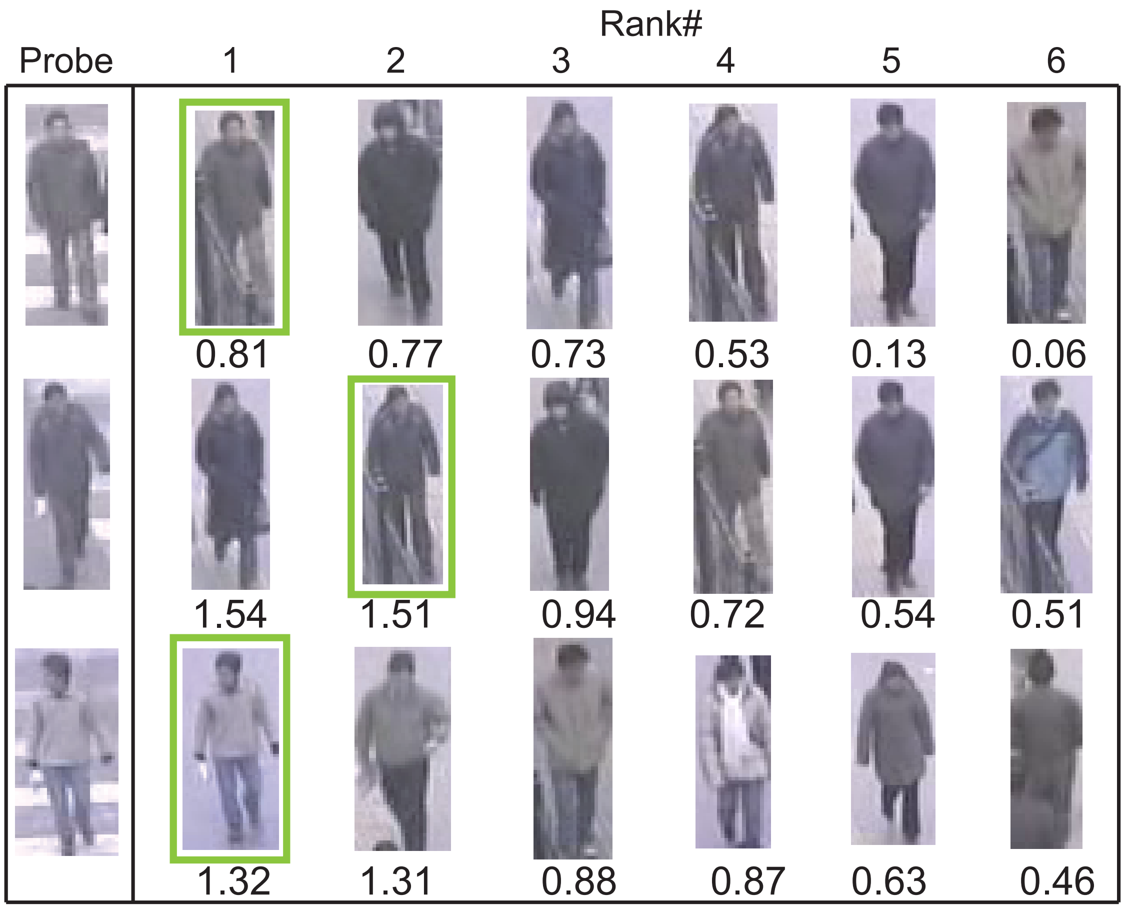


Figure 4.4: The top 6 matches of three example probes. The scores are shown below the images of matched persons. The ground-truth matches are bounded by green boxes.

4.3.2 Evaluation and Comparison

The CMC curve using our approach, which annotates as *Appearance+SAG+RF*, on NLPR_MCT dataset 1 and dataset 2 is shown in Figure 4.5 (a) and (b), respectively. The matches for rates at rank 1, 5, and 10 are given in Table 4.1. We show the top 6 matching of three example probes and the similarity scores in Figure 4.4.

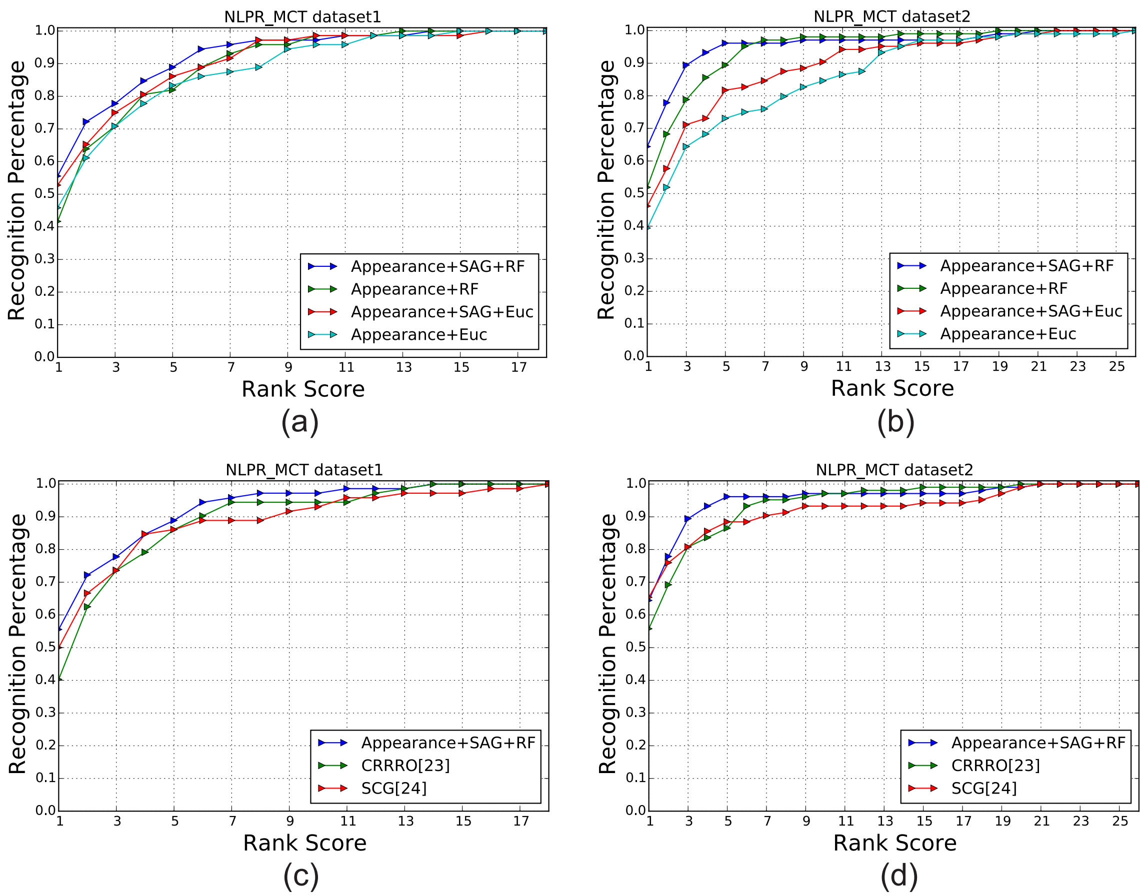


Figure 4.5: The CMC curve results. (a) and (b) are the CMC curve results using our approach (Appearance + SAG + RF) compare with Appearance-based approach (Appearance + RF) and metric-based approaches (Appearance + SAG + Euc, Appearance + Euc), on two datasets. (b) and (d) are the CMC curves using our approach compare with two state-of-the-art group-context based re-identification approaches (CRRRO[88] and SCG [77]).

To further investigate the contribution of group model, we compare our results to the re-identification results obtained by using appearance model only. We denote this as *Appearance+RF*. To investigate the improvement that is brought by random forest model, we also compare our results to the results obtained by using rather than using the score generated by random forest; we compute the Euclidean distance between features vectors as the matching score. We denote the approach that uses both appearance feature and SAG feature as *Appearance+SAG+Euc* and the approach that uses appearance feature only as *Appearance+Euc*.

The results are also shown in Figure 4.5 (a) and (b). The matching rate of all approaches at rank 1, 5 and 10 are shown in Table 4.1. As seen, our approach performs better than other approaches in both datasets, especially at the lower rank. By including the SAG feature as group information, the performance of both scoring approaches (random forest and Euclidian distance) is improved significantly. In dataset 1, *Appearance+SAG+RF* performs 14% and 7% better than *Appearance+RF* at rank 1 and 5, respectively. Although it performs slightly worse than other approaches at rank 10, it already reaches 97.2% accuracy. In dataset 2, *Appearance+SAG+RF* performs 12.5% and 6.7% better than *Appearance+RF* at rank 1 and 5. Similar to Dataset 1, it does not outperform other approaches at rank 10 but it reaches a high accuracy of 97.1%. This proves that the proposed SAG feature does capture the meaningful group information and reduces the ambiguity of re-identification task. By comparing the results of model-based approach (RF) and metric based approach (Euc), we can see that model-based matching also provides a the reasonable performance boost. In dataset 1, *Appearance+SAG+RF* performs 2.9% and 2.8% better than *Appearance+SAG+Euc* at rank 1 and 5, respectively. In dataset 2, *Appearance+SAG+RF* significantly out-performs *Appearance+SAG+Euc* by 18.3%, 14.4% and 6.8% at rank 1, 5 and 10, respectively.

Compare to state-of-the-art. We also compare our approach with Center Rectangular Ring Ratio-Occurrence (CRRRO) proposed by Zheng *et. al.* [88] and Subject Centric Group (SCG) feature proposed by Wei et. al. [77]. Since CRRRO feature can only be extracted from static group images, we create group images by cropping the video frames that include group members. The distance between CRRRO features are reverted and linearly combined with the scores that are generated by our *Appearance+RF* to form the final *CRRRO* scores. We extract SCG feature for each person in the dataset, the distance between SCGs are reverted to the similarity score of SCG features, which is combined with *Appearance+RF* score linearly to obtain the final *SCG* score.

The results are shown in Figure 4.5 (b) and (d) and the matching rate of approach *CRRRO* and *SCG* at rank 1, 5 and 10 are shown in Table 4.1. Comparing with *CRRRO*, our approach performs significantly better by 15.4% at rank 1 and 2.8% both at rank 5 and 10 in dataset 1. In dataset 2, our approach performs better by 8.7% and 9.6% at rank 1 and 5, respectively. Comparing with *SCG*, our approach out-performs *SCG* by 5.6%, 2.8% and 4.1% at rank 1, 5 and 10 in dataset 1, respectively In dataset 2, our approach performs identical with *SCG* at rank 1, and out-performs *SCG* by 7.7% and 3.9% at rank 5 and 10, respectively. In summary, our approach has the best overall performance in both datasets compared to twp other state-of-the-art methods.

4.4 Discussion

We address the problem of person re-identification using spatial appearance group feature. We introduce spatial appearance group feature that captures the group shape and group appearance from video frames. Our approach predicts the probability of two persons being the same when observed in different cameras using appearance model and group model. The appearance model is trained using the appearance feature. The group model is trained using proposed spatial appearance group feature. We demonstrate that our proposed method outperforms two other state-of-the-art group information based re-identification approaches. The limitation of the current approach is that it requires manually annotated data, which is very expensive to obtain. In the future, we plan to extend our method to learn the models using active learning, which requires much less amount of training data.

Chapter 5

Contextual Features For Human Activity Recognition

5.1 Introduction

Human activity analysis is one of the most important problems that has received considerable attention from the computer vision community in recent years. It has various applications, spanning from activity understanding for intelligence surveillance system to improving human-computer interactions. Recent approaches have demonstrated great performance in recognizing individual actions [81, 67]. However, in reality, human activity can involve multiple people. To recognize such group activities and their interactions would require information more than the motion of individuals. Therefore, human-activity recognition remains a challenging research topic largely due to the tremendous intra-class variation of human activities attributed to the visual appearance differences, subject motion variabilities, and viewpoint changes.

To solve these challenges, previous approaches in human activity recognition have focused on information about context. Context can be defined as information that is not directly related to the human activity itself, but it can be utilized to improve the traditional target-centered activity recognition [74]. Amer *et. al.* [44] proposes action context to encode the human interactions among multiple people. Choi *et. al.* [15] uses spatio-temporal volume descriptor to capture nearby person actions.

However, the existing approaches for human activity recognition mainly use people as context without richer context information, such as the scene information where the activity is performed, the location of the person within the scene, etc. Further, previous approaches have either utilized the context directly as feature inputs to classifiers such as random forest [15] and support vector machine [66], or incorporated context through probabilistic models like conditional random field [68]. There is little work utilizing deep models and networks to capture the contexts for human activity recognition. Deep models have the potential to systematically incorporate multiple sources of contexts due to their multi-level deep structure, the capability of probabilistic reasoning, and the integration of hidden units to synthesize higher level representations of the raw input features [74]. Therefore, in this work, we propose a deep-neural-network

(DNN) based model to recognize the human activity by taking advantage of its probabilistic reasoning power and incorporate multiple sources of context information. We combine motion and context information. The motion information is encoded by using the low-level motion features and high-level mobility features. The context information is incorporated to represent the scene and the human interactions. The scene feature encodes the attribute of the scene at the global and local level, while the group feature captures the human motion interaction and their spatial relationships in space. For each feature, we carefully design the network structure to get the higher level representation of input features, and the combination of different representations. We demonstrate that the integration of our context features and deep model can achieve better performance than state-of-the-art approaches on the collective activity dataset [14], which represent the human activities in real world scenario.

In summary, the main contributions of this work are:

- We introduce a two-level scene context descriptor. Beside the group context feature similar to many other works, we introduce a two-level scene context feature that describes the environment information of centered-target at the global and local levels.

- A deep model for human activity recognition. We present a deep neural network model that jointly captures multiple sources of context information, and achieves state-of-the-art performance over the collective activity dataset.

5.2 Method

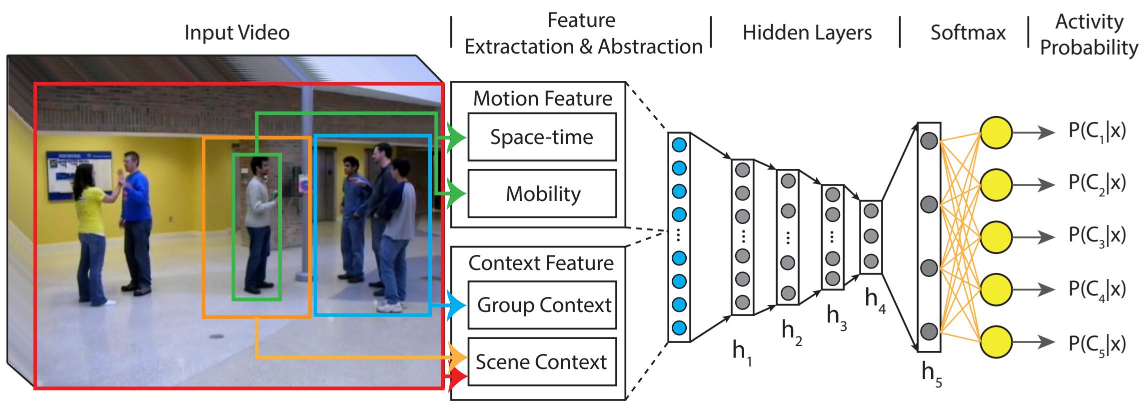


Figure 5.1: The structure of proposed neural network model for human activity recognition.

To recognize the human activity, we introduce a deep neural network with the structure of our network as shown in Figure 5.1.

Human activity recognition. Given the input video with tracking information of each subject, our system recognizes activity of each individual person at every frame. Two distinct features are considered in this recognition network, the first based on human motion (Sec. 5.2.1) and the second based on the context (Sec. 5.2.2). The features are extracted and abstracted using dense fully connected hidden layers. The output hidden units of the two parts are combined and fed into another fully connected network, which has a soft-max layer to compute the probability of recognized activity from input observations.

We use video dataset to train the deep model. In addition, we assume that the people tracking information, an estimation of their 3D space location, and facing direction in 3D space is available. We denote persons appearing in video as $\{p_i\}$, the tracking 2D bounding box of i-th person at frame t as b_i^t, the estimated 3D location as $l_i^t \in \mathbb{R}^3$, estimated the facing direction quantized into 8 viewpoints as $d_i^t \in \{front, frontleft, left, backleft, back, backright, right, frontright\}$.

In the following, we discuss in detail about our proposed features, along with our training and inferencing approach.

5.2.1 Motion Features

Motion features we consider are the low-level observation of the movements in the video. The introduced approach uses Space-time features that capture the low-level motion observed in the video and the mobility features that capture the movement of human as a whole part.

For an input video, we compute features for frames with interval β. That is, we extract features for the sample located at time t by computing the feature descriptors using a video segment comprising of frames in the interval $[t-\beta, t+\beta]$.

Space-time features. There are various space-time features to describe human motions in the video. We choose space-time interest points (STIP) [46], because it can extract feature points in space-time dimension robustly, and it also has been applied in event recognition [74]. STIP method detects interest points using a space-time extension of the Harris operator. For each interest point, it computes descriptors of the associated space-time patch. In this work, histograms of oriented gradient (HOG) and histograms of optical flow (HOF) feature are computed as the descriptors of the space-time patch. We obtain the feature words of both features by first detecting all the interest points over the entire videos data set, and then applying K-Means clustering to obtain K_i feature words for HOG features and HOF features.

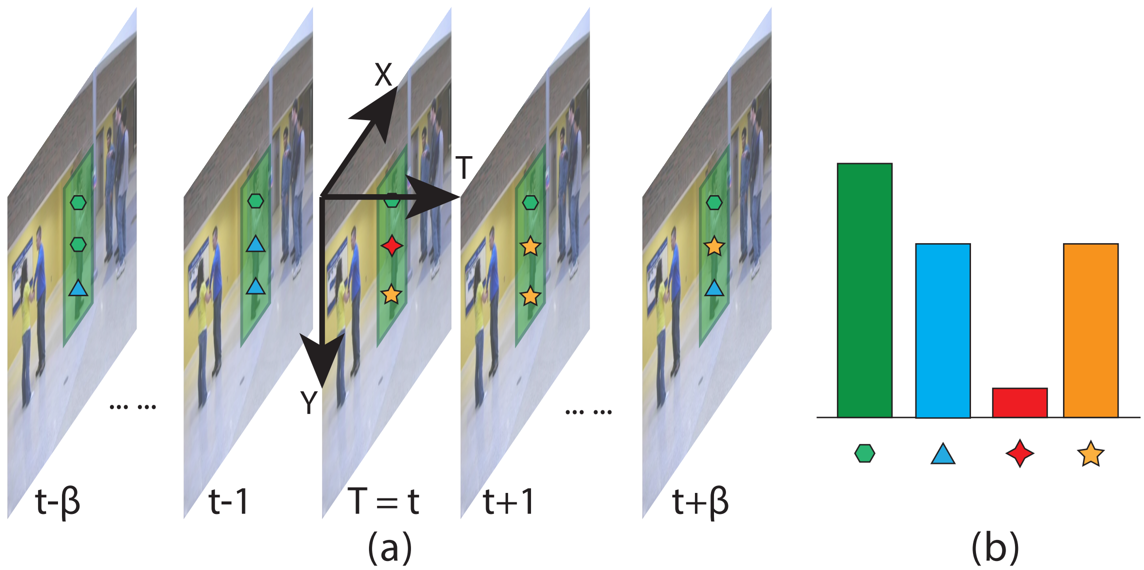


Figure 5.2: STIP feature histogram. (a) shows the video segment centered at time t, with length 2β+1. The green boxes denote the bounding box areas of the subject. (b) shows the STIP histogram generated using the video segment (a).

To describe the motion of p_i at time t, we first collect all the interest points located within $\{b_i^k | k \in [t-\beta, t+\beta]\}$, as shown in Figure 5.2(a). Then we compute the histogram of gradient and optical flow given the collected interest points as shown Figure 5.2(b). Finally, it results in two K_m dimensional histogram vectors.

After normalization to ensure each vector can sum up to one, the concatenation of two vectors serves as the motion descriptor of the person. We denote this as S_i^t. If there are no interest points located in bounding boxes of the subject, the descriptor is a zero vector of dimension $2K_m$. The extracted feature forms the input into our network as shown in the left part of Figure 5.3, and is followed by four fully-connected layers with $(h_{s1}, h_{s2}, h_{s3}, h_{s4})$ hidden units at each layer. Finally, at the top we have a layer with h_{s4} hidden units to realize a response to be combined with mobility information described below.

Mobility feature. As the estimated people 3D location can be computed using [20], we take the distance of movement in 3D space through the video segment as a description of human mobility. We compute subject movement at time t as $v_i^t = l_i^t - l_i^{t-1}$, where l_i^t denotes the location of p_i at time t. We denote p_i's mobility feature at time t is $V_i^t = [v_i^{t-\beta}, v_i^{t-\beta+1}, ..., v_i^t, ..., v_i^{t+\beta-1}, v_i^{t+\beta}]$, which is a vector of length $2\beta + 1$. We input the extracted mobility feature into our network as shown in the right part of Figure 5.3. The input layer is fully connected to a hidden layer that contains h_o units.

The hidden units of STIP features and mobility feature are concatenated to form a merge layer, which is fed into another fully connected layer of size h_m. These h_m hidden units abstract the overall motion information of the subject observed in the video at a sample frame.

5.2.2 Context Features

In our approach, context information plays an important role to improve the activity recognition accuracy. The context information includes two parts: the scene-based context and group-based context. Scene-based context captures the

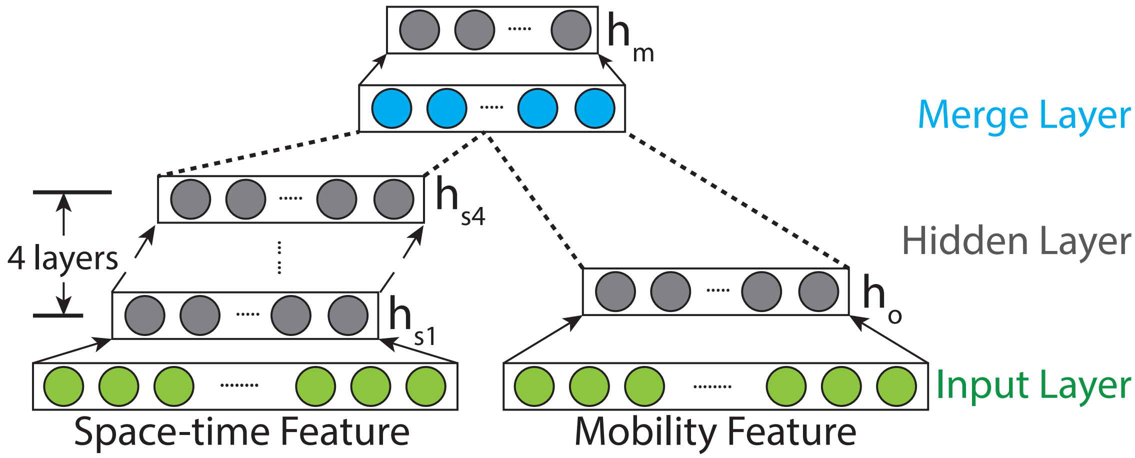


Figure 5.3: Motion feature layers. The green layers are network inputs; the gray layers are fully connected dense layers with hidden units; the blue layer is merge layer, which concatenates its inputs layers.

environment information surrounding the subject, allowing the network to find the association between environment information and activities. Scene-based context has two levels: **scene prior** describes the global scene attributes of the video frame; and **scene context** describes the scene around the person locally. The group of people that are physically near the subject also provide strong context information about human activity, as many approaches build various features to describe the people actions of near by humans to improve activity recognition accuracy [15, 66]. Similarly, the group-based context contains two parts: **group action** describes the interaction observation of nearby people; and **group structure** captures the shape (positions, direction) of nearby humans relative to the target person.

5.2.2.1 Scene-based context

Some activities have strong association with the environment, so the environment information as a context can reduce the ambiguity in its recognition. For example, jogging or crossing activities are more likely to happen in outdoor scenes, and queuing will be more likely to happen if the scene appears to be a shop. In this work, we extract the scene context information by looking into the image patch that extends around the bounding box of the tracked subject and use the descriptors of theses image patches as context features.

Rather than using low-level features such as appearance features to describe the image patches, we use a descriptor with semantic meaning. We utilize the existing place recognition methods to extract the semantic attribute of subject contexts. As deep convolution network gives the state-of-the-art performance in place recognition tasks, we use the Place-CNN [89] to generate the image patch descriptor. Given an image patch to Place-CNN, it outputs the probability of given image belonging to 205 categories. An example of place recognition on context image patches is shown in Figure 5.4. We simply denote the recognition process of Place-CNN as function $Place(I_t)$, where I_t is the image frame at time t of a given video, $Place(.)$ returns the probability vector of given image being recognized as belonging to the place categories.

Scene prior. The scene prior gives the environment context information at a global level for each video frame. To extract the scene prior feature, L^t, for all the

Figure 5.4: The scene prior and scene context. The green box is the bounding box of tracked people, with people id inside it. The yellow boxes are the scene context areas of persons. The red box which bounds the whole image is the scene prior area. We input images into Place-CNN to recognize place probability. The top two likely places of the above scene and scene context of person 1 and 2 are shown below the figure.

subjects that appear at time t, we compute

$$L^t = \frac{1}{2\beta + 1} \sum_{k=t-\beta}^{t+\beta} Place(I^k), \tag{5.1}$$

where scene prior feature $L^t \in \mathbb{R}^{205}$ and $\sum_{s=1}^{205} L_s^t = 1$.

Scene context. Besides the scene prior as global information for all the subjects appearing in the video frame, for each individual subject, we also build local scene features that capture the local environment information.

We denote the scene context image patch of p_i at time t as T_i^t, which is the region surrounding the bounding box b_i^t. Both T_i^t and b_i^t have the same center location, while width and height of T_i^t is 3 and 1.5 times the width and height of b_i^t, respectively. The scene context feature of p_i at time t is denoted as Q_i^t, which is computed as follow:

$$Q_i^t = \frac{1}{2\beta + 1} \sum_{k=t-\beta}^{t+\beta} Place(T_i^k), \tag{5.2}$$

Where $Q_i^t \in \mathbb{R}^{205}$ and $\sum_{s=1}^{205} Q_s^t = 1$.

After we compute scene-prior and scene-context features, we input the two features into the network as shown in Figure 5.5. We first concatenate the two features prior to feeding them to two fully connected layers h_{t1} and h_{t2}. The intent is to capture the interaction between global scene prior and local scene context. The hidden units in layer h_{t2} serve to provide the scene context information.

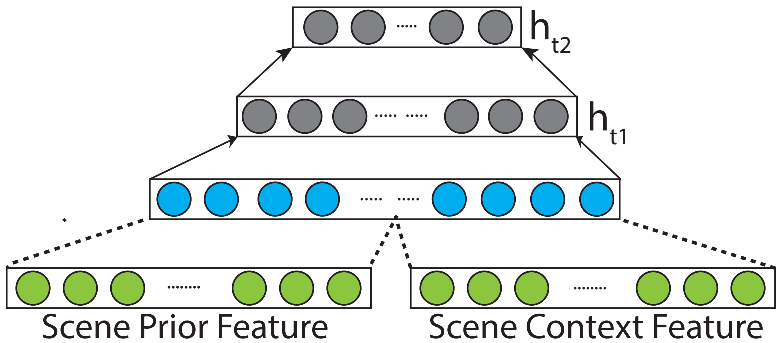


Figure 5.5: The network of combining scene prior and context information.

5.2.2.2 Group-based context

As people tend to form groups in various social behaviors, many approaches use the information from persons that physically are near the subject of interest to infer the activity. In our approach, we simply define the group as people within the social interaction area.

There are two group-based context information that are extracted: *group interaction context* captures the activity interaction of subject with group members; *group structure context* describe the spatial distribution of positions and directions of group members.

Group-interaction context. The group-interaction context captures the activity interactions between the centered subject and group members. We use concepts from proxemics [75, 66], and define interaction region as an area where the people are able to make social interaction with the centered subject. Interaction region is an ellipse $E(c_i, a, b)$, where the center of ellipse is c_i and (a, b) is the major and minor axis of ellipse, respectively. In our implementation, we use

$c_i = l_i + 0.3d_i$, $a = 3.35$, $b = 2.0$ as suggested in [75]. We are able to detect group members by finding the person within social interaction region, as shown in Figure 5.6. We denote the group members of subject p_i at time t as $N(p_i, t)$.

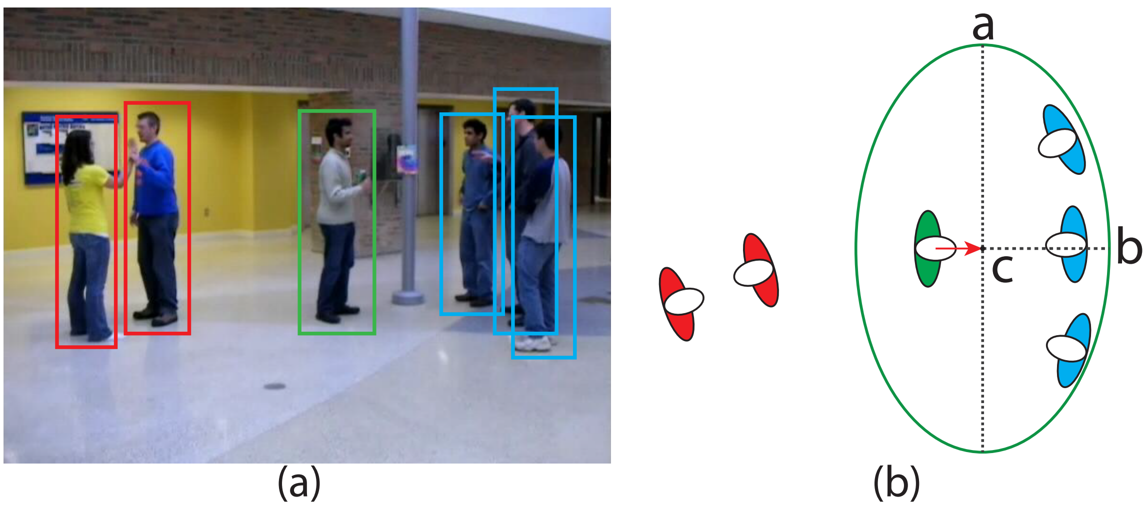


Figure 5.6: Interaction region. (a) The centered subject is in the green box, where the group members of target subject are in blue boxes, non-group members are in red boxes. (b) top view of persons 3D locations estimation of (a), the interaction region of the centered subject is displayed as the green ellipse, with center c and major a, minor b marked at ellipse.

To generate the group-interaction context feature for p_i at time t, we first compute the space-time features S_i^t, which is a bag-of-feature histogram of motion features as discussed in Sec. 5.2.1. Then we compute the average space-time feature U_i^t for all persons within the interaction region $N(p_i, t)$ as follow:

$$U_i^t = \frac{1}{|N(p_i, t)|} \sum_{p_j \in N(p_i, t)} S_j^t \tag{5.3}$$

We generate a 2D histogram as $B_i^t = {S_i^t}^{\top} * U_i^t$ that captures the co-occurrence frequencies of S_i^t and U_i^t. We normalize the 2D histogram B_i^t to ensure that all elements in the matrix sum to 1 and build a group interaction context feature by flattening the matrix into a K_i^2 dimension vector. If K_i is large, then we can re-create word bags for STIP features by clustering all the motion features of the data set. In our implementation we use $K_a = \sqrt{K_i}$ as number of bags for group interaction context feature extraction.

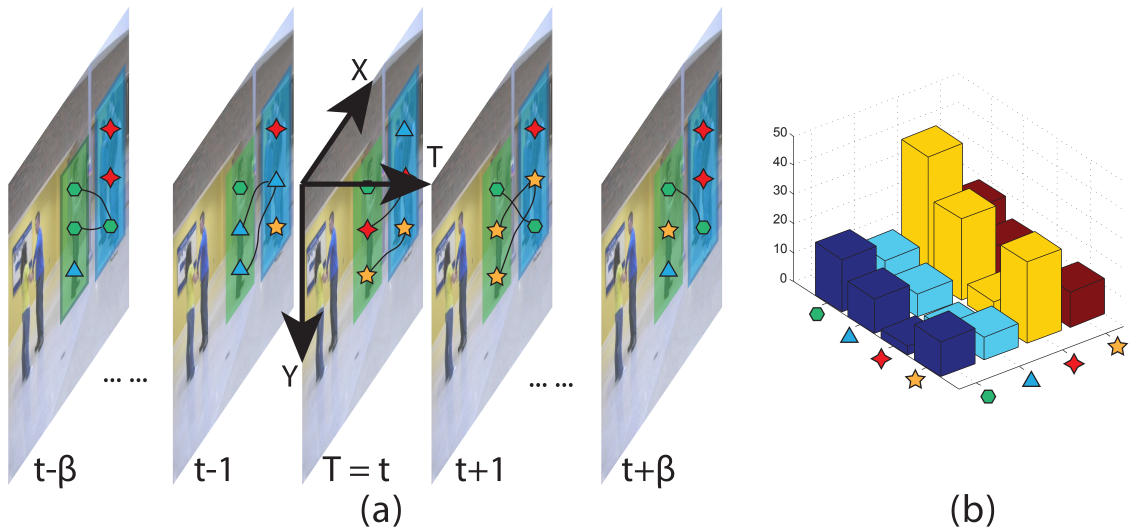


Figure 5.7: Group interaction context feature. (a) shows a video segment, where the green part covers the bounding box of target subject, the blue part covers the interacting group members. (b) shows the 2D co-occurrence histogram of target subject in the video segment (a).

Group-structure context. The group-structure context describes the relative positions and directions of people within interaction regions. For different activities, the shape of the group and the interactions between group members can be different. For example, group talking activity would have more than two people positioned in front of each other, face to face, while queuing activity most likely has more than two people standing in a line and facing the same direction. Therefore, we design group structure context feature to capture the positions and facing direction of the group.

To describe the position information, we construct a local coordinate centered at the target subject, as shown in Figure 5.8(a), and form a histogram of angles to represent the position distribution of group members. We denote the function

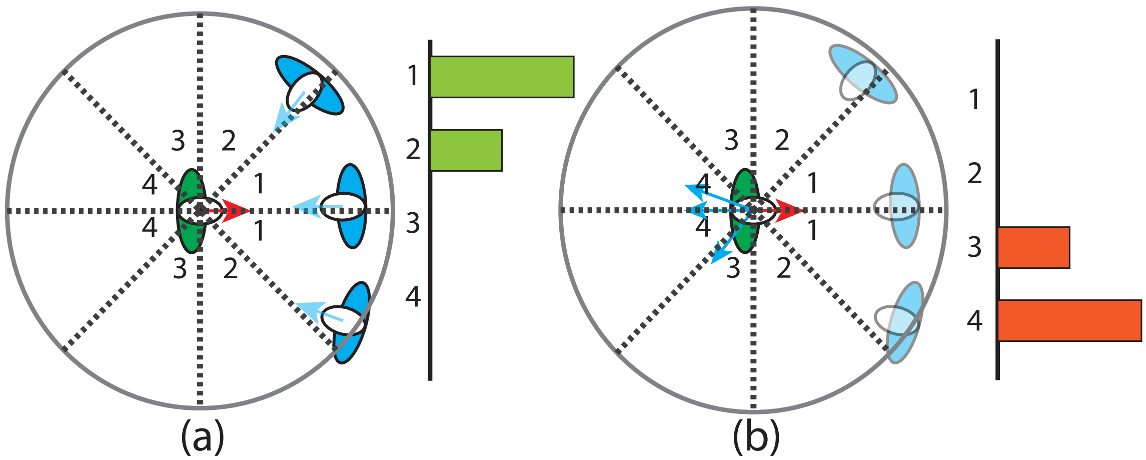


Figure 5.8: Group position histogram and direction histogram. (a) shows the position histogram; (b) shows the direction histogram. In this figure, the angle space is split into 4 sub-range in order to compute histogram.

$Ang(\mathbf{V_1}, \mathbf{V}_2)$ that returns the angles between vector $\mathbf{V_1}$ and $\mathbf{V_2}$. The group member position distribution of p_i at time t is obtain by computing the normalized histogram of angle set $\{Ang(d_i^k, l_j^k - l_i^k) | p_j \in N(p_i, k), k \in [t-\beta, t+\beta]\}$. To capture the direction information, we calculate the angles between the direction of centered subjects and other group members, then form a histogram of directions that represent the direction distribution of interacting neighbors, as shown in Figure 5.8(b). The group member direction distribution of p_i at time t is obtain by computing the normalized histogram of angle set $\{Ang(d_i^k, d_j^k) | p_j \in N(p_i, k), k \in [t-\beta, t+\beta]\}$. Both position and direction histogram have K_s bins.

After the position histogram and direction histograms are concatenated, we have the group structure context feature. We denote it as G_i^t, which is a $2K_s$ dimension feature, where K_s is bin size of angle histogram.

Finally, position histogram and direction histogram are input into two fully connected hidden layers (the hidden units number are h_{i1} and h_{i2} for group interaction feature; h_{r1} and h_{r2} for group structure feature), followed by a merge layer. The hidden units at the top represent the group context information.

After the final hidden layer for scene context information and group context information, we use a merge layer to concatenate hidden units from the two layers, as shown in Figure 5.9. The merge layer is fed into a fully connected layer for further abstraction. The top h_c hidden units form the representation for the overall context information of a given observation.

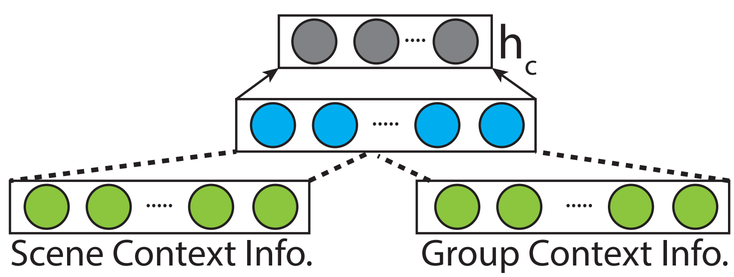


Figure 5.9: The network of group context informations.

The hidden units of motion information in Figure 5.3 and context information in Figure 5.9 are further concatenated, and input into the network shown in Figure 5.1, which includes four fully connected layers (with hidden units number h_1, h_2, h_3, h_4), and a soft-max layer at the end to calculate the probabilities of given observation for a set of activities. So far we have presented our deep neural network model, and in the following we are going to present the method for training and inference using our model.

5.2.3 Learning and Inference

Model Learning. The proposed model is a neural network with parameter W, which includes the weights matrix and bias parameters of all dense layers in the network. We denote $X = \{x_i^t, A_i^t | i = 1, ..., N, t = 1 + \beta, ...\}$ as the training data, where $x_i^t = (S_i^t, V_i^t, L^t, Q_i^t, B_i^t, G_i^t)$ includes all the individual features, and A_i is the ground truth human activity label. The output of the network is the probability of given observation belonging to each class of activity label. We denote the forward propagation as $F(W, x_i^t) = \{P(C_k | x_i^t), k = 1...M\}$, where M is the number of activity categories. In the training phase, we compute and minimize the categorical cross-entropy between predictions and ground truth:

$$E(W, X) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} 1(A_i = C_j) \log(P(C_j | x_i)) \tag{5.4}$$

We optimize the loss function using Stochastic Gradient Descent (SGD) updates with Nesterov momentum [59]. In each iteration, the model parameters W are updated as follow:

$$\triangle W_t = \mu * \triangle W_{t-1} - lr * \triangledown_W E(W_t + \mu * \delta W_{t-1}, Z) \tag{5.5}$$

$$W_{t+1} = W_t + \triangle W_t \tag{5.6}$$

Where μ is the momentum, lr is learning rate, $\triangledown_W$ is the gradient of the model parameter W, and Z is a random subset of training data for computing gradient in each iteration. We initialize the parameters of the network using Glorot weight initialization [25].

Model Inference. Given query human activity observation x, our model recognizes the activity category $C^{\star}$ by finding the maximal posterior probability given the observations from both motion feature space and context feature space through Equation 5.7.

$$C^{\star} = \arg\max_{k} P(C_k|x) \tag{5.7}$$

We implement our network using Lasagne [18] with GPU acceleration.

5.3 Experiments and Comparisons

In this section we describe the experiments that evaluate the performance of the proposed model for human activity recognition.

5.3.1 Evaluation Datasets

Our human activity recognition model is tested using Collective Activity dataset introduced by Choi et al. [14]. Other datasets (e.g. CAVIAR, VIRAT, or UT-Interaction) either focus on single person activity or the semantic labels provided are agnostic to scene context.

Collective Activity dataset comprises of two versions. The first version of data set contains five activities (*Cross*, *Wait*, *Queue*, *Walk* and *Talk*) and we denote this as Data-Act-5. The second version of dataset includes two additional activities (*Dance* and *Jog*) and removes the *Walk* activity, since the *Walk* activity is an individual activity rather than a collective activity. We denoted the second version as Data-Act-6. HOG based human detection and head pose estimation along with a probabilistic model is used to estimate camera parameters [14]. Extended Kalman filtering is employed to extract 3D trajectories and head pose estimates are provided as part of the dataset. In general, this dataset represents real-world, noisy observation with occlusions and automatic person detection and trajectory generation. We use the 4-fold cross-validation scheme similar to [13] to test the performance of our approach. To minimize the over-fitting in training phrase, we split data of non-training fold randomly into validation data set (30%) and testing data set (70%). In each interaction of parameter updates, the accuracy of validation data set is computed. When the accuracy over the training data set increases, but the accuracy over the validation data set stays the same or decreases, the neural network is over-fitting and we stop training.

5.3.2 Experiments and Comparison

In this section, we demonstrate the effectiveness of the proposed human activity recognition model that integrates both motion features and multiple sources of context information. The neural network to be evaluated has configuration as shown in Table 5.1.

The experiments are performed on both versions of Collective Activity dataset.

Table 5.1: Experiments Network Configuration

h_{s1}	150	h_m	25	h_{r1}	10	h_3	25
h_{s2}	100	h_{t1}	100	h_{r2}	10	h_4	25
h_{s3}	100	h_{t2}	20	h_c	10	h_5	25
h_{s4}	100	h_{i1}	10	h_1	50		
h_o	10	h_{i2}	10	h_2	50		

The performance of proposed model on both versions of the dataset is shown in Figure 5.10. The low value of the non-diagonal elements implies that our model is highly discriminative with low decision ambiguity between activities.

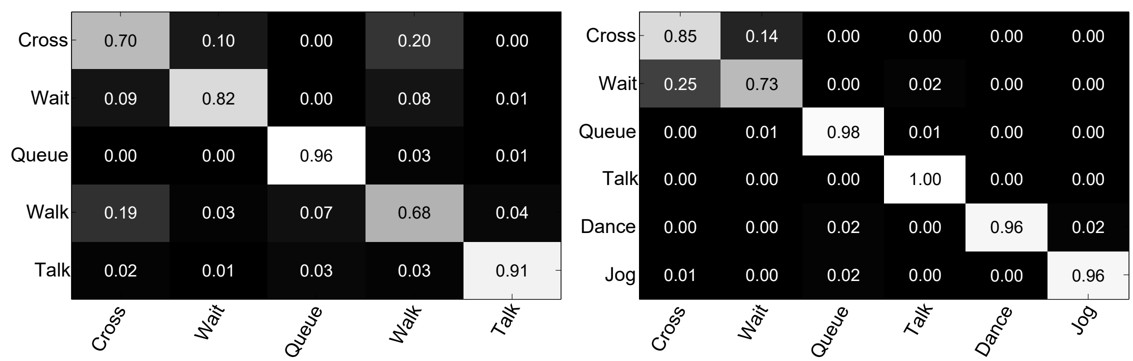


Figure 5.10: Confusion matrix of Collective Activity Dataset. 5 activities version (top) and 6 activities version (bottom).

The confusion matrix of Data-Act-5 at Figure 5.10(left) also shows that the confusion between *Walk* and *Cross* is reasonably low, despite the fact that both activities are *Walk* activity but with different scene semantics. Our model captures the scene context information and recognizes *Walk* activity better than baseline approaches as shown in Table 5.2 and other state-of-the-art approaches as shown in Table 5.3.

Compare with Baseline Approaches. To investigate the contribution of each

								Accuracy(%)
5-Activites	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(5 Act.)
SVM-Motion	36.2	64.8	52.0	28.3	20.5	-	-	40.4
DNN-Motion	46.4	63.9	49.1	44.0	53.6	-	-	51.4
SVM-Motion-Scene	42.5	65.2	65.7	43.3	51.0	-	-	53.5
DNN-Motion-Scene	49.5	64.7	65.3	54.8	84.6	-	-	64.8
SVM-Motion-Group	33.4	68.3	78.9	36.1	93.7	-	-	62.1
DNN-Motion-Group	39.3	74.0	97.4	78.8	96.0	-	-	77.1
SVM-Motion-Scene-Group	37.5	67.2	82.3	40.3	93.7	-	-	64.2
DNN-Motion-Scene-Group	67.6	70.2	96.2	81.6	91.5	-	-	81.4
6-Activites	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(6 Act.)
SVM-Motion	-	63.9	51.8	29.3	21.1	98.4	95.3	60.0
DNN-Motion	-	78.6	57.4	45.8	56.1	93.1	95.3	71.0
SVM-Motion-Scene	-	67.2	65.9	42.2	51.5	97.7	96.5	70.1
DNN-Motion-Scene	-	86.5	99.6	80.2	79.0	98.4	97.8	90.2
SVM-Motion-Group	-	75.2	79.1	46.9	88.9	99.9	99.7	81.6
DNN-Motion-Group	-	75.6	99.2	70.9	98.9	90.1	96.2	88.5
SVM-Motion-Scene-Group	-	83.1	81.9	50.6	93.4	99.9	99.7	84.8
DNN-Motion-Scene-Group	-	85.4	97.9	72.6	99.6	96.4	96.1	91.3

Table 5.2: Evaluation of Individual Components Contribution.

individual information that builds up the feature, we separate the features into three parts: *Motion* part denotes the space-time feature and mobility feature; *Scene* part denotes the scene prior and scene context feature; *Group* part denotes the group interaction feature and group structure context. We use the following combinations of above three parts **(Motion, Motion-Scene, Motion-Group, Motion-Scene-Group)** to train the deep neural network (*DNN*) model and compare their performance to validate the contribution of each individual part. When one part of feature is not involved in the training, we remove the nodes and layers related to that part within the network. To evaluate the discriminative power of proposed deep model, we take the same feature combinations and train the *Support Vector Machine (SVM)* classifier [12] and compare its performance with DNN model. We conduct the experiments on both Data-Act-5 and Data-Act-6 and the results are summarized in Table 5.2.

By looking into the average accuracy, **DNN-Motion-Scene** outperforms the **DNN-Motion** by 13.4% in Data-Act-5 and 19.2% in Data-Act-6, the activities that bring the significant accuracy improvements are *Talk*(31.0%) and *Queue*(16.2%) in Data-Act-5, *Queue*(42.3%) and *Wait*(34.4%) in Data-Act-6. **DNN-Motion-Group** outperforms the **DNN-Motion** by 25.7% in Data-Act-5 and 17.5% in Data-Act-5 in average. Interestingly, *Queue* and *Talk* provide the significant accuracy improvements in both datasets: *Queue* improves 48.4% in Data-Act-5 and 42.8% in Data-Act-5, *Talk* improves 42.4% in Data-Act-5 and 42.8% in Data-Act-6. The observed improvements are reasonable because queuing and talking activities have relatively stable group structures and interaction patterns, and these improvements indicate that our proposed group context feature captures the meaningful information for group structure and interaction. **DNN-Motion-Scene-Group** outperforms **DNN-Motion-Scene** by 17.9% in Data-Act-5 and 1.1% in Data-Act-6, it also outperforms **DNN-Motion-Group** by 4.3% in Data-Act-5 and 2.9% in Data-Act-6. This indicates that both scene context information and group context information contribute to the final performance improvements of the combined feature. However, the contribution rate of scene context information and group context information may vary among different datasets.

By comparing the accuracy of SVM classifier and deep neural network model that is trained using the same features, we are able to evaluate the discriminative power of the proposed deep model. Overall, the accuracy of DNN based model outperforms the SVM model by 13.3% in Data-Act-5 and by 11.1% in Data-Act-6. This clearly indicates that our proposed DNN model also contributes to higher performance of activity recognition task.

Compare with state-of-the-art. We also compare our results with other approaches that have state-of-the-art performance on Collective Activity dataset. For Data-Act-5, we compare our results with Spatio-Temporal Volume descriptor of Choi *et al.* [14] and Action Context descriptor of Lan *et al.* [44]. For Data-Act-6, the following methods are compared: the approach by Tran *et al.* [68] that uses group context descriptor, the approach by Amer *et al.* [2] that uses a chain model for group activities recognition and [3] that utilize top-down/bottom-up inference for activity recognition; and the approach by Choi *et al.* [15] that uses

									Accuracy(%)
Approaches	Year	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(5 Act.)
Choi et al. [14]	2009	57.9	55.4	63.3	64.6	83.6	-	-	65.9
Lan et al. [44]	2012	**68.0**	65.0	96.0	68.0	**99.0**	-	-	79.1
Our Method (5 Act.)		67.6	**70.2**	**96.2**	**81.6**	91.5	-	-	**81.4**
Approaches	Year	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(6 Act.)
Choi et al. [15]	2011	-	76.5	78.5	78.5	84.1	94.1	80.5	82.0
Amer et al. [2]	2011	-	69.9	96.8	74.1	**99.8**	87.6	70.2	83.1
Amer et al. [3]	2012	-	77.2	95.4	78.3	98.4	89.4	72.3	85.1
Khai et al. [68]	2015	-	60.6	89.1	**80.9**	93.1	93.4	95.4	85.4
Our Method (6 Act.)		-	**85.4**	**97.9**	72.6	99.6	**96.4**	**96.1**	**91.3**

Table 5.3: Comparison with state-of-the-art approaches.

random forest for activities recognition.

The results are shown in Table 5.3. We can see that our approach performs best in 3 out of 5 activities in Data-Act-5, and 4 out of 6 activities in Data-Act-6. Our approach also gives the best average accuracy for both datasets. Finally, our approach outperforms other approaches by 2% in Data-Act-5 and at least by 5% in Data-Act-6.

5.4 Discussion

In conclusion, this chapter proposes a deep-neural-network model for human activity recognition from video. The input features of the deep network include motion feature and context feature. We design the scene prior feature and scene context feature to capture the environment around the subject of interest global and local levels. We demonstrate that our model can outperform state-of-the-art human activity recognition methods in the collective activities dataset.

Chapter 6

Conclusion

This thesis presents multiple human groups contextual feature and introduces techniques that use these features in video analysis application, including human re-identification and human activity recognition. The rich experiments have demonstrated the proposed approaches successfully on a variety of datasets that represent real life scenarios. This section summarizes the contributions of this thesis to the field of human re-identification and human activity recognition and presents an overview of future work.

6.1 Summary of Key Contributions

The main contributions of this thesis are human group contextual descriptors along with algorithms for human re-identification and human activity recognition in videos.

This thesis introduces Subject-Centric Group feature and Spatial-Appearance Group Feature for the person re-identification problem. Subject-Centric Group (SCG) feature can capture the positions and appearance information of group members walking aside a centric subject. By using group-shape-based and group-appearance-based metrics, the distance between subject centric group features can be calculated. The SCG feature provides an unsupervised and easily implemented approach that can improve the human re-identification accuracy of existing individual-based human re-identification. Spatial-Appearance Group (SAG) feature is a fixed length descriptor that encodes the group appearance and structure around a subject. Machine learning model can be trained to learn the appearance and group structure changing across the camera. This supervised algorithm allowing the user to use more training data and improve the performance of introduced group matching model. Combined with appearance based machine learning model, we demonstrate the improvements that been brought using proposed feature and models.

In the field of human activity recognition, this thesis presents contextual features to capture context information for recognizing human activities. Group interaction features are introduced to capture the interaction between people in the group. In specific, group direction histogram and group position histogram is introduced to capture the spatial relationship between human and the belonging group.

The environment context information is discovered to be useful in reducing the ambiguity of human activity in video. Environment context is designed in two scales: at the global scale, the scene prior feature is used to describe the environment of entire video; at the local scale, scene context feature is used to encode the environment of a subject. The environment context feature takes advantage of fast growing deep learning technologies and uses place-CNN to generate both environments contextual features. To train a machine learning model by considering multiple features jointly, we introduce a deep model that take all the features as input, and output the probabilities of giving feature is generated by the certain activity. To recognize the activity of input feature, they firstly will go through two to three fully connected layer to obtain abstract representations of the features. Then these abstract representations are combined and pass through several additional hidden layers to calculate the probabilities. The features and algorithm are evaluated using Collective Behavior Dataset and introduced method can reach state-of-the-art performance.

6.2 Limitations and Future Work

The are several limitations using the group as context information in person re-identification application. Firstly, introduced algorithm needs to assume a stable group structure across the cameras, which means the cameras have to be within a reasonable distance so that group appears in one camera will show in the other camera without changing the group members. This assumption will be challenged in case the scale of camera network is large. The chance of group structure changing is higher as cameras distance become longer. Although we also introduce the appearance-based metric for subject-centric group feature, it still not robust in handling the change of group members. Therefore, the future works can consider the relationship between group members and non-group members to infer the group structure changing to provide more accurate matching among the individuals appeared in the video. Secondly, both introduced re-identification approach is built based on other human re-identification features. However, as it demonstrated in subject centric group feature work, the improvement of introduced context feature is bounded by the robustness of baseline appearance feature. Further works should look consider directly calculate the group features from group image and videos without associating with high-level features of an individual. Thirdly, introduced approach both focused on person re-identification for two people appeared in two cameras. However, most security camera networks have a much larger scale. People appears in multiple cameras would contain more contextual information that useful to obtain the identity. Therefore, the research about how to extend current work to larger camera network is also a topic could be explored in the future.

For human activity recognition works, there are also some limitations in current works. Firstly, the current activity recognition approach is performed frame-by-frame without considering the temporal coherence. Further works can utilize a different machine learning model that considering temporal information, such as Hidden Markov Model and recurrent networks, assuming training datasets of larger scale are available. Secondly, the current works only recognized the human activities that are manually annotated in the training dataset. This fact limits the application of introduced approach because the currently available human activity datasets have very limited action category (Collective Activity Dataset has only 6 categories). Furthermore, human being often performing multiple activities at the same time. For example, one person can walk and talk with another person at the same time, while in current datasets, the person is only labeled walking. As the deep neural network is growing dramatically in recent years, we already see very promising applications of DNN in areas like image-based object recognition and natural language processing. It will be interesting to see how the state-of-the-art deep networks models can provide richer context information in human activity recognition area.

Bibliography

[1] Multi-camera object tracking challenge, http://mct.idealtest.org, Augest 2014.

[2] M. Amer and S. Todorovic. A chains model for localizing participants of group activities in videos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 786–793, Nov 2011.

[3] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Computer Vision–ECCV 2012*, pages 187–200. Springer, 2012.

[4] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1619–1632, 2011.

[5] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *Image Processing, IEEE Transactions on*, 20(6):1709–1724, 2011.

[6] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[7] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram. Human activity recognition using multidimensional indexing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(8):1091–1104, 2002.

[8] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

[9] I. Biederman, J. C. Rabinowitz, A. L. Glass, and E. W. Stacy. On the information extracted from a glance at a scene. *Journal of experimental psychology*, 103(3):597, 1974.

[10] Y. Cai and G. Medioni. Exploring context information for inter-camera multiple target tracking. *IEEE Winter Conference on Applications of Computer Vision*, pages 761–768, Mar. 2014.

[11] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Computer Vision-ECCV 2004*, pages 350–362. Springer, 2004.

[12] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[13] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision–ECCV 2012*, pages 215–230. Springer, 2012.

[14] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282–1289, Sept 2009.

[15] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[16] J. L. Davenport and M. C. Potter. Scene consistency in object and background perception. *Psychological Science*, 15(8):559–564, 2004.

[17] P. De Graef, A. De Troy, and G. D'Ydewalle. Local and global contextual constraints on the identification of objects in scenes. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 46(3):489, 1992.

[18] S. Dieleman, J. Schlter, C. Raffel, E. Olson, S. K. Snderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. D. Fauw, M. Heilman, diogo149, B. McFee, H. Weideman, takacsg84, peterderivaz, Jon, instagibbs, D. K. Rasul, CongLiu, Britefury, and J. Degrave. Lasagne: First release., Aug. 2015.

[19] N. Endo and Y. Takeda. Use of spatial context is restricted by relative position in implicit learning. *Psychonomic bulletin & review*, 12(5):880–885, 2005.

[20] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[21] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[22] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[23] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):1003–1016, 2012.

[24] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1528–1535. IEEE, 2006.

[25] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

[26] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales. The re-identification challenge. In *Person Re-Identification*, pages 1–20. Springer, 2014.

[27] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. 2014.

[28] P. Grother and P. J. Phillips. Models of large population recognition performance. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–68. IEEE, 2004.

[29] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–6. IEEE, 2008.

[30] A. Hampapur, L. Brown, J. Connell, S. Pankanti, A. Senior, and Y. Tian. Smart surveillance: applications, technologies and implications. *Information, Communications and Signal Processing*, 2:1133–1138, 2003.

[31] J. Han and B. Bhanu. Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2):316–322, 2006.

[32] M. Hasan and A. K. Roy-Chowdhury. Continuous learning of human activity models using deep nets. In *Computer Vision–ECCV 2014*, pages 705–720. Springer, 2014.

[33] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*. Springer, 2011.

[34] H. S. Hock, G. P. Gordon, and R. Whitehurst. Contextual relations: the influence of familiarity, physical plausibility, and belongingness. *Perception & Psychophysics*, 16(1):4–8, 1974.

[35] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, 2013.

[36] K. Jia and D.-Y. Yeung. Human action recognition using local spatio-temporal discriminant embedding. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[37] S. Jörg, J. Hodgins, and A. Safonova. Data-driven finger motion synthesis for gesturing characters. *ACM Transactions on Graphics (TOG)*, 31(6):189, 2012.

[38] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014.

[39] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi. A review on video-based human activity recognition. *Computers*, 2(2):88–131, 2013.

[40] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[41] H. Keval. Cctv control room collaboration and communication: Does it work? In *Proceedings of Human Centred Technology Workshop*, pages 11–12, 2006.

[42] T. Ko. A survey on behavior analysis in video surveillance for homeland security applications. In *Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE*, pages 1–8. IEEE, 2008.

[43] S. Kumari and S. K. Mitra. Human action recognition using dft. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 Third National Conference on*, pages 239–242. IEEE, 2011.

[44] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(8):1549–1562, 2012.

[45] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[46] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[47] K. H. Lee, M. G. Choi, Q. Hong, and J. Lee. Group behavior from video: a data-driven approach to crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 109–118. Eurographics Association, 2007.

[48] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. In *ACM Transactions on Graphics (TOG)*, volume 29, page 124. ACM, 2010.

[49] Y. Li, Z. Wu, and R. J. Radke. Multi-shot re-identification with random-projection-based random forests. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 373–380. IEEE, 2015.

[50] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[51] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[52] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 413–422. Springer, 2012.

[53] C. Madden, E. D. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18(3-4):233–247, 2007.

[54] O. Marques, E. Barenholtz, and V. Charvillat. Context modeling in computer vision: techniques, implications, and applications. *Multimedia Tools and Applications*, 51(1):303–339, 2011.

[55] R. Martín-Félez and T. Xiang. Gait recognition by ranking. In *Computer Vision–ECCV 2012*, pages 328–341. Springer, 2012.

[56] C. McPhail and R. T. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Methods & Research*, 10(3):347–375, 1982.

[57] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, pages 69–82. Springer, 2004.

[58] G. Nebehay and R. Pflugfelder. Clustering of Static-Adaptive correspondences for deformable object tracking. In *Computer Vision and Pattern Recognition*. IEEE, June 2015.

[59] Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Papers*, (2007/76), 2007.

[60] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, pages 28–39. Springer, 2004.

[61] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. BMVC*, pages 21.1–11, 2010. doi:10.5244/C.24.21.

[62] A. Rabinovich and S. Belongie. Scenes vs. objects: a comparative study of two approaches to context based recognition. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 92–99. IEEE, 2009.

[63] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, pages 1–20, 2000.

[64] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(2):162–177, 2005.

[65] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36 Vol.3, Aug 2004.

[66] K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah. Social cues in group formation and local interactions for collective activity analysis. In *VISAPP*, pages 539–548, 2013.

[67] K. N. Tran, I. A. Kakadiaris, and S. K. Shah. Part-based motion descriptor image for human action recognition. *Pattern Recognition*, 45(7):2562–2572, 2012.

[68] K. N. Tran, X. Yan, I. A. Kakadiaris, and S. K. Shah. A group contextual model for activity recognition in crowded scenes. In *VISAPP*, 2015.

[69] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008.

[70] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

[71] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013.

[72] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1505–1518, 2003.

[73] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *Computer Vision–ECCV 2014*, pages 688–703. Springer, 2014.

[74] X. Wang and Q. Ji. Video event recognition with deep hierarchical context model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4418–4427, 2015.

[75] J. Was, B. Gudowski, and P. J. Matuszyk. Social distances model of pedestrian dynamics. In *Cellular Automata*, pages 492–501. Springer, 2006.

[76] L. Wei and Z. Deng. A practical model for live speech-driven lip-sync. *Computer Graphics and Applications, IEEE*, 35(2):70–78, 2015.

[77] L. Wei and S. K. Shah. Subject centric group feature for person re-identification. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, pages 28–35, June 2015.

[78] L. Wei and S. K. Shah. Person re-identification with spatial appearance group feature. In *Technologies for Homeland Security (HST), 2016 IEEE Symposium on*, pages 1–6. IEEE, 2016.

[79] L. Wei and S. K. Shah. Human activity recognition using deep neural network with contextual information. In *International Conference on Computer Vision Theory and Applications*, 2017.

[80] L. Wei, W. Yu, M. Li, and X. Li. A non-rigid registration algorithm for compatible skeletonization. In *Computer Science and Education (ICCSE), 2010 5th International Conference on*, pages 209–214. IEEE, 2010.

[81] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

[82] D. Williams. Effective cctv and the challenge of constructing legitimate suspicion using remote visual images. *Journal of Investigative Psychology and Offender Profiling*, 4(2):97–107, 2007.

[83] F. Xu and M. Gao. Human detection and tracking based on hog and particle filter. In *Image and Signal Processing (CISP), 2010 3rd International Congress on*, volume 3, pages 1503–1507. IEEE, 2010.

[84] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *Computer Vision–ECCV 2014*, pages 536–551. Springer, 2014.

[85] A. Yilmaz, X. Li, and M. Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1531–1536, 2004.

[86] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2528–2535. IEEE, 2013.

[87] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 144–151. IEEE, 2014.

[88] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proc. BMVC*, pages 23.1–23.11, 2009. doi:10.5244/C.23.23.

[89] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.